Reverse-Engineering the Reader

Samuel Kiegeland[*1] **Ethan Gotlieb Wilcox**[*1] **Afra Amini**[1]
David Robert Reich[2,3] **Ryan Cotterell**[1]
[1]ETH Zürich [2]University of Potsdam [3]University of Zürich
skiegeland@ethz.ch {ethan.wilcox, afra.amini, ryan.cotterell}@inf.ethz.ch
reich@cl.uzh.ch

Abstract

Numerous previous studies have sought to determine to what extent language models, pretrained on natural language text, can serve as useful models of human cognition. In this paper, we are interested in the opposite question: whether we can directly optimize a language model to be a useful cognitive model by aligning it to human psychometric data. To achieve this, we introduce a novel alignment technique in which we fine-tune a language model to implicitly optimize the parameters of a linear regressor that directly predicts humans' reading times of in-context linguistic units, e.g., phonemes, morphemes, or words, using surprisal estimates derived from the language model. Using words as a test case, we evaluate our technique across multiple model sizes and datasets and find that it improves language models' psychometric predictive power. However, we find an inverse relationship between psychometric power and a model's performance on downstream NLP tasks as well as its perplexity on held-out test data. While this latter trend has been observed before (Oh et al., 2022; Shain et al., 2024), we are the first to induce it by manipulating a model's alignment to psychometric data.

1 Introduction

Language comprehension is thought to be predictive and incremental. Research on reaction times (Fischler and Bloom, 1979), fixation patterns (Ehrlich and Rayner, 1981), and brain activations (Kutas and Hillyard, 1984; DeLong et al., 2005) suggests that comprehenders anticipate upcoming linguistic units based on the context in which they occur (Kuperberg and Jaeger, 2016).[1] In addition, a large body of evidence shows that when linguistic units are unexpected, they require more cognitive effort to process (Miller and McKean, 1964; Ehrlich and Rayner, 1981; Balota et al., 1985, *inter alia*). E.g., reading times of units, e.g., morphemes, words, and sentences, are taken as a measure of cognitive effort, i.e., the less likely the unit is in context, the longer it takes to read (Smith and Levy, 2013).

In this study, we are interested in reverse-engineering the part of the language processing system responsible for predicting abstract linguistic units and testing it by measuring its ability to predict reading times.[2] To do so, we need, first, to establish a theoretical link between predictability and reading times. For this, we draw on **surprisal theory** (Hale, 2001; Levy, 2008), which posits that the cognitive effort to process a unit is proportional to its surprisal—the unit's negative log probability given the preceding context. Implicitly, surprisal theory assumes that a comprehender maintains a probability distribution over upcoming units, i.e., it assumes a *human* language model. However, this human language model is a theoretical construct, and cannot be observed directly. Thus, most previous work that tests surprisal theory has done so using probability estimates derived from a language model trained on large swathes of human-written text. Under this paradigm, it has been observed that surprisal estimates derived from language models, fit with regularized maximum-likelihood estimation on large corpora, do yield significant predictors although these findings vary based on the quality (Goodkind and Bicknell, 2018; Wilcox et al., 2020, 2023b), size, and training-data set size (Oh et al., 2022; Shain et al., 2024) of the model. One current line of research therefore seeks to uncover what characteristics of pretrained LMs produce better predictors of human reading times, and what this tells us about the human language processing system.

Our paper asks a simple question: Instead of assessing the ability of pretrained LMs to serve as psycholinguistic predictors, can we directly esti-

[*] Equal contribution.

[1] Our code is available at https://github.com/samuki/reverse-engineering-the-reader.

[2] In our experiments, we predict abstract units that correspond to words. However, in order to frame our discussion more generally and to align it with our mathematical framework, we use the term **units** instead.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 9367–9389
November 12-16, 2024 ©2024 Association for Computational Linguistics

mate (or fine-tune) a language model so that its surprisal estimates become better predictors of processing effort for linguistic units? We frame this problem as one of aligning the language model to human data (Christiano et al., 2017; Schulman et al., 2017; Ouyang et al., 2022; Ziegler et al., 2020; Rafailov et al., 2023). However, in contrast to much previous work, which uses human–model alignment to obtain improvement on natural language processing tasks, e.g., summarizing text (Stiennon et al., 2020) or producing non-toxic outputs (Li et al., 2024), we seek to align LMs to be better psychometric predictors. Particularly, we aim to directly align models to human *reading* data by optimizing the parameters of the statistical models typically used to evaluate the psychometric fit. While recent approaches like direct preference optimization (DPO; Rafailov et al., 2023) are designed to optimize a model's parameters based on human preferences, they rely on pairwise preference data, which is not applicable to real-valued psychometric data. Thus, we propose a novel alignment technique that allows us to directly optimize the language model's parameters in such a way that it serves as a better predictor of real-valued psychometric data. Specifically, we fine-tune the language model to implicitly optimize the coefficients of a linear regression that predicts the reading time of an individual unit.

We test our technique on three English-language reading datasets and find that it increases the statistical fit of a linear regressor in terms of the likelihood it assigns to reading times on a held-out test set. We also observe a positive relationship between a model's psychological predictive power and its perplexity. While it has been observed that better LMs are better psychological models of reading up to a point (Goodkind and Bicknell, 2018; Wilcox et al., 2020, 2023a), after a certain model size, their fit to human reading times decreases (Oh et al., 2022; Shain et al., 2024). In other words, beyond an inflection point, better LMs are *worse* predictors of human reading times. Through our alignment procedure, we are able to demonstrate the contrapositive, namely that as we causally make our language models' outputs more aligned with reading, they become worse at predicting the next word.

2 Psycholinguistics Background

Put concisely, the goal of this paper is to reverse-engineer p_{H}, a person's internal language model from psychometric data collected through experimentation. Our reasoning is as follows: if we can align an existing language model $p_{\boldsymbol{\theta}}$ to more accurately predict such psychometric data, $p_{\boldsymbol{\theta}}$ will also more closely resemble p_{H}.

2.1 Language Models

Let Σ be an **alphabet**, i.e., a finite, non-empty set, and let $\overline{\Sigma} \overset{\text{def}}{=} \Sigma \cup \{\text{EOS}\}$ be the alphabet augmented with a distinguished end-of-string symbol not in Σ. A **language model** $p_{\boldsymbol{\theta}}$ is a probability distribution over Σ^*, which is the set of all strings over Σ. Further, following Opedal et al. (2024), we define the **normalized prefix probability**

$$\pi_{\boldsymbol{\theta}}(\boldsymbol{c}) \overset{\text{def}}{=} \frac{1}{\mathrm{Z}_{\pi_{\boldsymbol{\theta}}}} \sum_{\boldsymbol{u} \in \Sigma^*} p_{\boldsymbol{\theta}}(\boldsymbol{c}\boldsymbol{u}), \tag{1}$$

which is a probability distribution over prefixes $\boldsymbol{c} \in \Sigma^*$, where $\boldsymbol{cu}$ is the concatenation of $\boldsymbol{c}$ and $\boldsymbol{u}$. The normalization constant $\mathrm{Z}_{\pi_{\boldsymbol{\theta}}} = 1 + \sum_{\boldsymbol{u} \in \Sigma^*} p_{\boldsymbol{\theta}}(\boldsymbol{u})|\boldsymbol{u}|$ ensures that all probabilities sum to one. Here $|\boldsymbol{u}|$ denotes the length, i.e., the number of units in a string $\boldsymbol{u}$. Note that Eq. (1) is only well-defined in an LM with finite expected length.

2.2 Psychometric Measurements

Let $\psi(u, \boldsymbol{c}) \in \mathbb{R}$ denote a measurement for a unit $u \in \overline{\Sigma}$ appearing in context $\boldsymbol{c} \in \Sigma^*$. In this paper, $\psi(u, \boldsymbol{c})$ represents various reading time measurements for a given unit, such as gaze duration, first fixation duration, and total fixation duration, which are standard approximations to the processing effort of a linguistic unit in context (Miller and McKean, 1964; Just and Carpenter, 1980; Frazier and Rayner, 1982; Rayner, 1998, *inter alia*).

2.3 Surprisal Theory

Surprisal theory furnishes us with an easy-to-compute predictor of processing effort that is derived from a pretrained language model. Formally, surprisal theory predicts that the time it takes to process a linguistic unit $u \in \overline{\Sigma}$ in context $\boldsymbol{c} \in \Sigma^*$ is an affine[3] function of the unit's contextual surprisal under the human language model p_{H}, defined as

$$\iota_{\mathrm{H}}(u \mid \boldsymbol{c}) \overset{\text{def}}{=} -\log_2 p_{\mathrm{H}}(u \mid \boldsymbol{c}). \tag{2}$$

Surprisal theory has been supported by numerous empirical studies, which have found that surprisal is predictive of reading times across multiple datasets

[3] Previous work has often described the relationship as linear. However, we use affine here due to the additive constant required to link the two.

(Smith and Levy, 2013; Wilcox et al., 2020; Shain et al., 2024), types of reading time measurements (Pimentel et al., 2023), and languages (Kuribayashi et al., 2021; Meister et al., 2021; Wilcox et al., 2023b). Typically, p_{H} is approximated using a LM estimated from corpus data, i.e., we substitute

$$\iota_{\boldsymbol{\theta}}(u \mid \boldsymbol{c}) \overset{\text{def}}{=} -\log_2 p_{\boldsymbol{\theta}}(u \mid \boldsymbol{c}) \tag{3}$$

in for ι_{H} when predicting processing effort.

2.4 Linear Modeling

We now discuss how empirical support for surprisal theory is typically adduced. Following previous work, we assume an affine function links a linguistic unit's contextual surprisal and that unit's reading time[4] (Smith and Levy, 2013; Wilcox et al., 2023b; Shain et al., 2024) and apply linear regression to predict reading times based on contextual surprisal. In mathematical jargon, both the psychometric measurements $\psi(u, \boldsymbol{c})\colon \overline{\Sigma} \times \Sigma^* \to \mathbb{R}$ and our predictors $\mathbf{x}_{\boldsymbol{\theta}}(u, \boldsymbol{c})\colon \overline{\Sigma} \times \Sigma^* \to \mathbb{R}^D$ are real-valued random variables. In the case of the predictor, given a unit $u \in \overline{\Sigma}$ and a context $\boldsymbol{c} \in \Sigma^*$, we define the predictor as a D-dimensional real column vector

$$\mathbf{x}_{\boldsymbol{\theta}}(u, \boldsymbol{c}) = [\iota_{\boldsymbol{\theta}}(u \mid \boldsymbol{c}), x^2, \ldots, x^D]^\top, \tag{4}$$

which, as depicted, includes our surprisal estimate $\iota_{\boldsymbol{\theta}}(u \mid \boldsymbol{c})$; the additional variables $x^2, \ldots, x^D$ are considered to be baseline predictors and are chosen at the modeler's discretion depending on what they seek to test. Given a parameter (column) vector $\boldsymbol{\beta}_{\boldsymbol{\theta}} \in \mathbb{R}^D$, we define the following linear model

$$\psi(u, \boldsymbol{c}) \sim f_{\boldsymbol{\beta}_{\boldsymbol{\theta}}}(\cdot \mid \mathbf{x}_{\boldsymbol{\theta}}(u, \boldsymbol{c})) \tag{5a}$$
$$= \mathcal{N}(\widehat{\psi}_{\boldsymbol{\beta}_{\boldsymbol{\theta}}}(u, \boldsymbol{c}), \sigma^2), \tag{5b}$$

where the linear function

$$\widehat{\psi}_{\boldsymbol{\beta}_{\boldsymbol{\theta}}}(u, \boldsymbol{c}) = \mathbf{x}_{\boldsymbol{\theta}}(u, \boldsymbol{c})^\top \boldsymbol{\beta}_{\boldsymbol{\theta}} \tag{6}$$

constitutes the mean and σ^2 is the variance.

We evaluate the predictive power of our model by fitting $f_{\boldsymbol{\beta}_{\boldsymbol{\theta}}}$ on a training set and measuring the log-likelihood of the test set; higher log-likelihood indicates greater predictive power of the model. To assess how much surprisal contributes to the predictive power, we fit two regression models based on two predictors. The baseline predictor $\mathbf{x}_{\mathrm{b}}(u, \boldsymbol{c})$, defined identically to Eq. (4), but with the estimated surprisal zeroed out, typically consists of a unit's unigram surprisal, i.e., its negative log unigram probability[5] and a unit's length (in characters). The target predictor, denoted by $\mathbf{x}_{\boldsymbol{\theta}}(u, \boldsymbol{c})$, includes the same set of baseline predictors together with the estimated surprisal of the unit u.

To quantify the predictive power of contextual surprisal, we compute the delta log-likelihood Δ_{llh} between the two models, which is the average unit-level difference in log-likelihood assigned by the two predictors to the reading time measurements. For a single unit–context pair, we compute:

$$\begin{aligned}\Delta_{\mathrm{llh}}(u, \boldsymbol{c}) = \log f_{\boldsymbol{\beta}_{\boldsymbol{\theta}}}(\psi(u, \boldsymbol{c}) \mid \mathbf{x}_{\boldsymbol{\theta}}(u, \boldsymbol{c})) \\ - \log f_{\boldsymbol{\beta}_{\mathrm{b}}}(\psi(u, \boldsymbol{c}) \mid \mathbf{x}_{\mathrm{b}}(u, \boldsymbol{c})),\end{aligned} \tag{7}$$

where $\boldsymbol{\beta}_{\mathrm{b}}$ and $\boldsymbol{\beta}_{\boldsymbol{\theta}}$ are the coefficients for the baseline and target models, respectively, and are estimated separately. Intuitively, a higher Δ_{llh} indicates that the estimated surprisals contribute more to the predictive power or psychometric accuracy of the model over reading times, compared to the baseline predictors (Frank and Bod, 2011).

Having established a metric, delta log-likelihood, to measure how much contextual surprisal contributes to predicting reading times, we are now ready to answer the question we posed at the beginning: Can we fine-tune a language model such that surprisal estimates derived from it become better predictors of reading times?

3 Aligning LMs to Psychometric Data

As discussed in §2.4, the psychometric predictive power of a language model is typically evaluated by assessing the predictive power of surprisal estimates $\iota_{\boldsymbol{\theta}}(u \mid \boldsymbol{c})$ of a linguistic unit u in a context $\boldsymbol{c}$ with respect to human reading times. Rather than *evaluating* a language model's psychometric predictive power, in this study, we ask whether we can fine-tune language models to *increase* their psychometric predictive power.

We treat this as an alignment problem (Christiano et al., 2017; Schulman et al., 2017; Ouyang et al., 2022; Ziegler et al., 2020; Rafailov et al., 2023). Let p_{ref} denote a pretrained language model that will serve as a reference. Further, let $p_{\boldsymbol{\theta}}$ denote the language model we seek to fine-tune, generally initialized to p_{ref}. Our goal is to align $p_{\boldsymbol{\theta}}$ such that its surprisal estimates are more directly correlated with psychometric data in comparison

[4] Although, see Hoover et al. (2023) for a different perspective on the shape of the linking function.

[5] See Opedal et al. (2024) for a detailed explanation of this predictor variable.

to p_{ref}. By means of implicit differentiation, we derive an objective that allows us to perform such psychometric alignment.

3.1 Deriving an Objective

Reward Function. We draw inspiration from direct preference optimization, which implicitly optimizes the parameters of a reward model for predicting the outcomes of pairwise comparisons between items. However, rather than implicitly fitting a Bradley–Terry model (Bradley and Terry, 1952) to model pairwise preferences given by human annotators, we implicitly fit a linear regressor $f_{\boldsymbol{\beta}_{\boldsymbol{\theta}}}$ to model psychometric measurements. Following the notation developed in §2.4, let $\mathbf{x}_{\boldsymbol{\theta}}(u, \boldsymbol{c})$ denote the predictor vector, which includes the contextual surprisal derived from the model $p_{\boldsymbol{\theta}}$. To optimize the parameters of $f_{\boldsymbol{\beta}_{\boldsymbol{\theta}}}$, we define our reward function as the *negative*[6] minimum of the expected mean squared error (MSE) between the observed psychometric data $\psi(u, \boldsymbol{c})$ and the predicted values $\widehat{\psi}_{\boldsymbol{\beta}_{\boldsymbol{\theta}}}(u, \boldsymbol{c}) = \mathbf{x}_{\boldsymbol{\theta}}(u, \boldsymbol{c})^\top \boldsymbol{\beta}_{\boldsymbol{\theta}}$ as defined in Eq. (6). The reward is then given by

$$\mathrm{r}(\boldsymbol{\theta}) \overset{\text{def}}{=} - \min_{\boldsymbol{\beta}_{\boldsymbol{\theta}} \in \mathbb{R}^D} \mathop{\mathbb{E}}_{(u,\boldsymbol{c}) \sim \pi_{\text{ref}}} \left(\psi(u, \boldsymbol{c}) - \widehat{\psi}_{\boldsymbol{\beta}_{\boldsymbol{\theta}}}(u, \boldsymbol{c})\right)^2, \tag{8}$$

where π_{ref}, as defined in Eq. (1), is the normalized prefix probability. Note that under our linear model specified in Eq. (5), maximizing $\mathrm{r}(\boldsymbol{\theta})$ is equivalent to maximizing the Δ_{llh} in Eq. (7).[7]

Regularization with KL Divergence. To prevent the fine-tuned model $p_{\boldsymbol{\theta}}$ from diverging excessively from the pretrained reference model p_{ref}, we regularize our objective with the Kullback–Leibler (KL) divergence, as is typically done in RLHF (Schulman et al., 2017; Ziegler et al., 2020; Stiennon et al., 2020; Ouyang et al., 2022). More specifically, the regularization term is defined as:

$$\varphi(\boldsymbol{\theta}) \overset{\text{def}}{=} \mathop{\mathbb{E}}_{\boldsymbol{c} \sim \pi_{\text{ref}}} \mathrm{KL}\big(p_{\text{ref}}(\cdot \mid \boldsymbol{c}) \mid\mid p_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{c})\big) \tag{9a}$$

$$= \mathop{\mathbb{E}}_{\boldsymbol{c} \sim \pi_{\text{ref}}} \sum_{u \in \overline{\Sigma}} p_{\text{ref}}(u \mid \boldsymbol{c}) \log \frac{p_{\text{ref}}(u \mid \boldsymbol{c})}{p_{\boldsymbol{\theta}}(u \mid \boldsymbol{c})}. \tag{9b}$$

where π_{ref} is the normalized prefix probability of the reference distribution p_{ref}.

Putting it All Together. We now combine the reward and the KL regularization to define an objective for aligning LMs to psychometric data as

$$\mathcal{J}(\boldsymbol{\theta}) \overset{\text{def}}{=} \underbrace{\mathrm{r}(\boldsymbol{\theta})}_{\text{reward}} - \underbrace{\lambda \cdot \varphi(\boldsymbol{\theta})}_{\text{KL reg.}}, \tag{10}$$

where $\lambda \geq 0$ is a hyperparameter, which determines the strength of the KL regularization. Because optimizing $\mathrm{r}(\boldsymbol{\theta})$ corresponds to optimizing Δ_{llh}, $\mathcal{J}(\boldsymbol{\theta})$ trades off better alignment with human psychometric data against the KL divergence from the pretrained model p_{ref}.

3.2 Approximation of the Reward Function

In practice, we use a Monte Carlo estimate of N unit–context pairs $(u_n, \boldsymbol{c}_n) \sim \pi_{\text{ref}}$ to approximate the expectation in Eq. (8). Let $\boldsymbol{\psi} = [\psi(u_1, \boldsymbol{c}_1), \ldots, \psi(u_N, \boldsymbol{c}_N)]^\top \in \mathbb{R}^N$ denote the real column vector of N reading time observations. Then we define the approximate reward as

$$\widetilde{\mathrm{r}}(\boldsymbol{\theta}) \overset{\text{def}}{=} - \min_{\boldsymbol{\beta}_{\boldsymbol{\theta}} \in \mathbb{R}^D} \frac{1}{N} \sum_{n=1}^{N} \left(\psi(u_n, \boldsymbol{c}_n) - \widehat{\psi}_{\boldsymbol{\beta}_{\boldsymbol{\theta}}}(u_n, \boldsymbol{c}_n)\right)^2 \tag{11a}$$

$$= - \min_{\boldsymbol{\beta}_{\boldsymbol{\theta}} \in \mathbb{R}^D} \frac{1}{N} ||\boldsymbol{\psi} - \mathbf{X}_{\boldsymbol{\theta}} \boldsymbol{\beta}_{\boldsymbol{\theta}}||^2, \tag{11b}$$

where $\mathbf{X}_{\boldsymbol{\theta}}$ is an $N \times D$ real matrix, with each row corresponding to a predictor vector, as defined in Eq. (4). Leveraging a well-known closed-form solution (see App. A.1), we directly compute Eq. (11b). To ensure that $\mathbf{X}_{\boldsymbol{\theta}}^\top \mathbf{X}_{\boldsymbol{\theta}}$ is invertible, we add a small regularization term $\rho \mathbf{I}$ with $\rho > 0$, leading to the following coefficients

$$\boldsymbol{\beta}_{\boldsymbol{\theta}}^\star = \big(\underbrace{\mathbf{X}_{\boldsymbol{\theta}}^\top \mathbf{X}_{\boldsymbol{\theta}} + \rho \mathbf{I}}_{\text{always invertible}}\big)^{-1} \mathbf{X}_{\boldsymbol{\theta}}^\top \boldsymbol{\psi}. \tag{12}$$

This results in the simple reward term:

$$\widetilde{\mathrm{r}}(\boldsymbol{\theta}) = -\frac{1}{N} ||\boldsymbol{\psi} - \mathbf{X}_{\boldsymbol{\theta}} \boldsymbol{\beta}_{\boldsymbol{\theta}}^\star||^2. \tag{13}$$

Notably, the optimal coefficients are now parameterized by the language model's parameters $\boldsymbol{\theta}$.[8]

4 Experimental Design

In this section, we discuss how we experiment with fine-tuning a language model using the objective defined in Eq. (10) and the reward approximation

[6] We use a negative sign to ensure that maximizing the reward corresponds to minimizing the prediction error.

[7] This is equivalent to maximum likelihood estimation. We omit $f_{\boldsymbol{\beta}_{\text{b}}}$ since it does not depend on $\boldsymbol{\theta}$.

[8] To compute the optimal coefficients $\boldsymbol{\beta}_{\boldsymbol{\theta}}^\star$ in Eq. (12) efficiently, we use the Cholesky decomposition; see App. A.2.

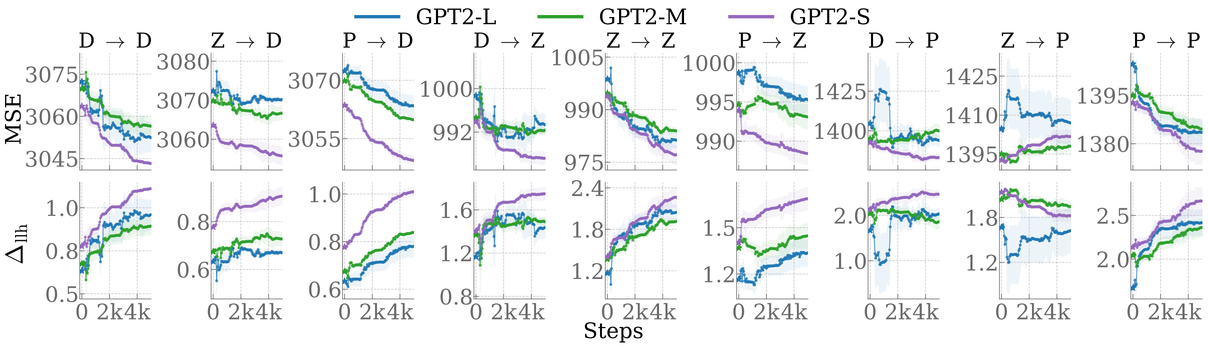


Figure 1: Learning curves for the MSE (top) and Δ_{llh} (10^{-2} nats, bottom) on the test datasets throughout fine-tuning. Bands show the standard error across random seeds. MSE tends to decrease, while Δ_{llh} increases, showing better prediction of reading times.

given in Eq. (13). Specifically, we design experiments to evaluate the effectiveness of our objective for improving a model's psychometric accuracy, i.e., how well it predicts human reading times. Additionally, we assess the impact of our objective on a model's quality, as measured through its perplexity on test data and its performance on downstream NLP tasks.

4.1 Models

We use the GPT-2 family of models (Radford et al., 2019) and conduct experiments on the `small`, `medium`, and `large` versions of the model available on the HuggingFace hub (Wolf et al., 2020).

4.2 Data

While our objective given in Eq. (10) assumes unit–context pairs sampled from p_{ref}, we lack psychometric data for LM-generated text. Instead, we use text from existing eye-tracking datasets, which we take to be a reasonable approximation, given p_{ref} assigns a high probability to the respective texts. However, future work should investigate this assumption more thoroughly. We fine-tune and evaluate models on three widely used eye-tracking corpora: The **Dundee Corpus** (Kennedy et al., 2003), which includes eye-movement data from 10 English-speaking participants reading 2368 sentences of newspaper articles from *The Independent*, the **Provo Corpus** (Luke and Christianson, 2018), which contains eye-tracking data from 84 participants who read 55 paragraphs of texts from various sources, including fiction and non-fiction, and the **ZuCo Corpora** (ZuCo 1.0 (Hollenstein et al., 2018) and ZuCo 2.0 (Hollenstein et al., 2020)), which contain data from 12 and 18 participants, respectively, reading sentences from Wikipedia articles and movie reviews.

Similar to previous work (Wilcox et al., 2020, 2023b), we focus on the gaze duration, defined as the total time of a reader's first pass fixations on a unit u before they fixate on a different unit; see App. B.1 for details. In addition, we conduct experiments using the total reading duration and first fixation duration (App. D.1). We further verify that our results are not due to random effects through additional experiments with random reading times (App. D.2). We create test sets by sampling 40% of the data from each corpus. Then, to construct various test sets, we randomly sample 70% of the remaining 60% of the data according to 3 random seeds. We can view this procedure as a simple bootstrapping procedure, from which we can approximate error bars (Efron, 1979). We fine-tune and evaluate models on all pairs of eye-tracking corpora, resulting in 9 unique data splits as shown in Tab. 4.

4.3 Fine-Tuning

We compute the contextual surprisal of each unit in a sentence, excluding units with zero reading times and zero frequencies, during fine-tuning. It is common practice in psycholinguistics literature to drop words that were skipped on the first pass during reading and, therefore, have a first pass reading time of zero (Smith and Levy, 2013; Oh and Schuler, 2023b). Across all experiments, the predictor $\mathbf{X}_{\boldsymbol{\theta}}$ consists of a unit's contextual surprisal, its unigram surprisal, estimated using Speer's (2022) toolkit, and its length. We opt not to include coefficients for spillover. We do so because reading times in eye-tracking studies have been observed to be less susceptible to spillover than other reading modalities, for example, self-paced reading (Shain

Configuration	Train Size	Test Size
Dundee (D) → Provo (P)	20894.7	1114
Dundee (D) → Dundee (D)	20894.7	20207
Dundee (D) → ZuCo (Z)	20894.7	7715
ZuCo (Z) → Provo (P)	7761	1114
ZuCo (Z) → Dundee (D)	7761	20207
ZuCo (Z) → ZuCo (Z)	7761	7715
Provo (P) → Provo (P)	1113.7	1114
Provo (P) → Dundee (D)	1113.7	20207
Provo (P) → ZuCo (Z)	1113.7	7715

Table 1: Data splits and configurations for fine-tuning and evaluation. Numbers indicate the mean number of tokens in each split across our random seeds.

Model	GPT2-L	GPT2-M	GPT2-S
D → D	55.291	32.822	42.880
P → D	23.294	24.233	29.898
Z → D	9.084	11.077	16.598
D → P	31.356	4.928	17.780
P → P	45.273	15.621	24.769
Z → P	2.116	6.460	1.999
D → Z	36.577	11.950	24.502
P → Z	15.012	5.730	20.094
Z → Z	81.390	41.801	61.196

Table 2: Percentage increase between the initial and maximum Δ_{llh} for each model. For exact start and maximum Δ_{llh} values, see Tab. 7.

and Schuler, 2021). Although, we acknowledge that this is a limitation of our study. For all configurations, we fine-tune models for 5k steps and repeat each run using three different random seeds. For an overview of all hyperparameters, see App. B.3.

4.4 Evaluation

We evaluate the predictive power of the estimated surprisal values for predicting reading times every 50 steps using our test data splits. During each evaluation phase, we first obtain surprisal estimates on the test data from the aligned language model. Using these estimates, we perform a 5-fold cross-validation on the test data, where we fit baseline and target linear regressors using ordinary least squares and use them to compute the unit-level mean Δ_{llh}.

5 Results

We now return to our main question: Does our objective align language models more closely to human reading times compared to pretrained models? We first analyze the effect of maximizing the unregularized reward and later, in §5.4, extend our analysis to the KL-regularized objective.

5.1 Predicting Reading Times

Are fine-tuned models $p_{\boldsymbol{\theta}}$ better predictors of reading times compared to pretrained models p_{ref}? To answer this question, we examine both the mean square error (MSE) and the delta log-likelihood Δ_{llh} computed at each evaluation step on the test data using cross-validation. While the Δ_{llh} is our main metric for measuring a model's psychometric accuracy, we use the MSE to compare the magnitude of prediction errors. The MSE is computed similarly to the unregularized objective, given in Eq. (10), except we calculate the MSE through cross-validation on the entire test set using linear regression to evaluate the predictive power of surprisal estimates.

As visualized in Fig. 1, the MSE values are relatively high because our reading times are in milliseconds; an MSE of 4,000 corresponds to a prediction that is off by only about 1/20$^{\text{th}}$ of a second. We observe that the MSE decreases for all models across all held-out datasets over the course of fine-tuning. An exception to this is the MSE for the data splits Dundee → Provo and ZuCo → Provo, where we do not observe consistent decreases in MSE, potentially due to the small size of the test set. Models evaluated on the ZuCo dataset have the lowest MSE, followed by Provo and Dundee. Further, in Fig. 1, we visualize the Δ_{llh} of regressors evaluated on our test dataset over the course of fine-tuning. In line with our observed decreases in MSE, we find that Δ_{llh} increases on most data splits, with the exception of models fine-tuned on Dundee or ZuCo and evaluated on Provo. Smaller models start with higher Δ_{llh}, which is consistent with previous literature (Oh and Schuler, 2023b). In Tab. 2, we compare the percentage increase from each model's start Δ_{llh} to the maximum Δ_{llh} it achieves over the course of the fine-tuning, averaged over three random seeds. Interestingly, GPT2-S shows higher percentage increases compared to GPT2-M, suggesting that model size alone does not fully account for these differences.

5.2 Coefficient Estimates

Additionally, we want to know what the fine-tuned models have implicitly learned about the role of surprisal, as well as the baseline features, over the course of fine-tuning. To examine this, in Fig. 3, we visualize the regressor coefficients $\boldsymbol{\beta}_{\boldsymbol{\theta}}$ from cross-validation on the whole test set. We observe the following tendencies: The coefficients for a unit's contextual surprisal and length are positive,

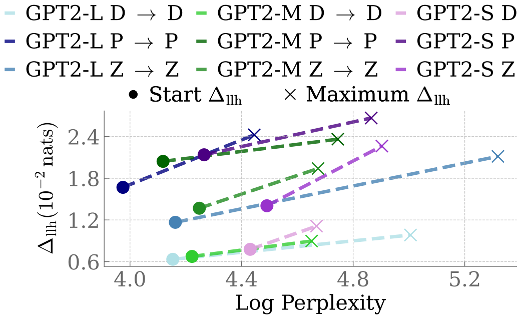


Figure 2: **Perplexity vs. Δ_{llh}.** We compare model perplexity at the start of fine-tuning to the point where they achieve the highest mean Δ_{llh}. Optimizing models using $\mathcal{J}$ increases perplexity. See Fig. 6 for all data splits.

indicating that, as units become less predictable and longer, they take more time to read. The positive coefficient for unigram surprisal means more frequent units take less time to read. These results align with the previous literature, e.g., those from Wilcox et al., 2023b. Interestingly, we observe that over the course of fine-tuning, the coefficient for surprisal tends to *increase*, while the coefficient for length *decreases*. For unigram surprisal and our bias term, i.e., the intercept of the regression, we observe mixed results, with coefficients for unigram surprisal decreasing for models evaluated on the Dundee and ZuCo datasets and increasing for models evaluated on Provo. Overall, the coefficient's trajectories suggest that the predictive power of surprisal for predicting reading times increases throughout fine-tuning. For coefficients across data splits, as well as coefficients for the regularized objective, see App. E.

5.3 Perplexity vs. Quality

Several recent papers have found that, above a certain size, as a language model's perplexity decreases, its predictive power increases (Oh et al., 2022; Shain et al., 2024). Follow-up work has suggested that this is due to the super-human predictive abilities of the models, especially for low-frequency nouns such as named entities (Oh and Schuler, 2023b). An open question remains regarding the causal factors behind these dynamics. So far, studies have tested the relationship by manipulating a language model's quality either by choosing different sizes of pretrained models as in Oh and Schuler (2023b) or by training successively smaller and smaller models as in Wilcox et al. (2023a). However, our fine-tuning methods allow us to flip the causal arrow. As we make a language model more closely aligned with human reading times, what happens to its quality? To investigate this question, in Fig. 2, we show the perplexity and Δ_{llh} both at the start of fine-tuning and when the model achieves its maximum Δ_{llh}, which is typically near the end of fine-tuning. We find that increases in Δ_{llh} generally correspond to increases in perplexity. These results indicate that as we make a language model's predictions more aligned with reading times, it becomes worse at modeling text.

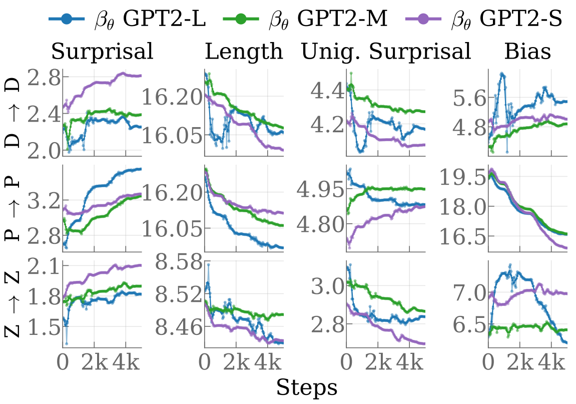


Figure 3: **Mean Coefficients of unit-level features over fine-tuning.** Smoothed values (window size 5) are shown, with unsmoothed values in a pale version of the color. Coefficients corresponding to surprisal tend to increase over the course of fine-tuning.

5.4 Kullback–Leibler Regularization

Next, we analyze the effect of adding KL regularization to our objective in Eq. (10). Specifically, we compare the trajectories of Δ_{llh}, KL divergence and log perplexity for KL coefficients $\lambda \in \{0, 5, 50, 500\}$. The coefficients used in this paper are larger than the ones normally used in RLHF studies (Schulman et al., 2017; Ziegler et al., 2020; Stiennon et al., 2020; Ouyang et al., 2022) because of the large magnitude of our reward in Eq. (13), which is the negative mean squared error. Fig. 4 shows a clear trend: Higher coefficients lead to lower increases in perplexity and KL divergence, as well as lower increases in Δ_{llh}. We observe higher coefficients reduce the divergence of $p_{\boldsymbol{\theta}}$ from their initial distribution p_{ref}. While higher coefficient come at the cost of lower Δ_{llh} increases, we still observe consistent increases over the baseline; see Fig. 9 for results on all data splits with $\lambda = 500$.

6 Additional Analyses

Previous work has found that language models fine-tuned on cognitive data, such as eye tracking (Yang and Hollenstein, 2023; Deng et al., 2023) and brain

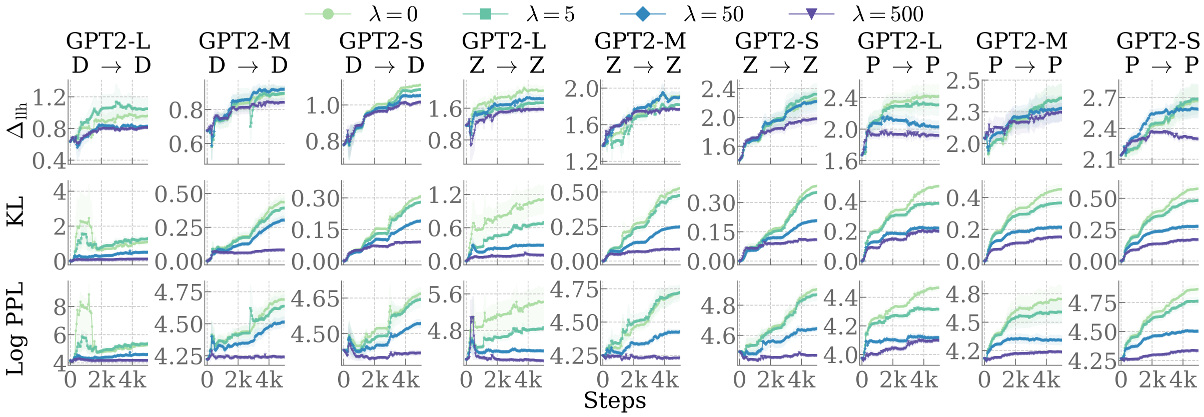


Figure 4: **Trajectories of Δ_{llh}, KL divergence and log perplexity for KL coefficients $\lambda \in \{0, 5, 50, 500\}$.** Higher coefficients lead to lower perplexity increases as well as lower Δ_{llh} increases, showing that the KL regularization constrains $p_{\boldsymbol{\theta}}$ from diverging too much from p_{ref}.

data (Toneva and Wehbe, 2019) can improve performance on downstream NLP tasks. In this section, we evaluate our fine-tuned language models—based on the lowest loss on the test dataset—on several such tasks. We observe no such improvement.

6.1 Targeted Syntactic Evaluation

We evaluate models on the Benchmark of Linguistically Minimal Pairs (BLiMP; Warstadt et al., 2020).[9] BLiMP assesses whether language models' behavior is consistent with human preferences for grammatical sentences across a range of grammatical phenomena. In BLiMP, items come in grammatical and ungrammatical variants; we report model accuracy for assigning a higher probability to the grammatical version. Models' scores on BLiMP are visualized in Fig. 5. Even though KL regularization helps mitigate the drop in accuracy, we observe that our fine-tuned models generally exhibit slightly lower accuracy than their non-fine-tuned counterparts, indicating that our fine-tuning procedure does not lead to a better generalization about English grammatical rules in our models.

6.2 Text Generation

Additionally, we analyze how our objective affects the ability of fine-tuned models to generate text and focus on two aspects: the uniformity of information and the diversity of the generations. To assess uniformity, we draw on the uniform information density (UID) hypothesis (Fenk and Fenk, 1980; Levy and Jaeger, 2006), which posits that language users prefer information to be evenly distributed throughout an utterance. A recent study by Meister et al. (2021) provides empirical support for the UID hypothesis in naturally occurring corpora and shows a link between linguistic acceptability and information uniformity. Here, we ask whether aligning models to human reading times encourages them to generate text with greater uniformity of information.

We test this by generating completions for prefixes from the CNN/DailyMail dataset (Hermann et al., 2015; See et al., 2017). In Tab. 3, we report the mean surprisal variance (UID$_v$) and unique unigram ratio (1-Gram%) across all completions; see App. F for more details. Models without regularization ($\lambda = 0$) show higher surprisal variance compared to the pretrained models, indicating less uniformly distributed information in the generations. However, under regularization ($\lambda = 500$), this trend is reversed, and we observe lower variance with the exception of GPT2-L: Dundee → Dundee. We further observe a decrease in the unique unigram ratio, indicating that fine-tuned models generate more repetitive text. Overall, these findings suggest that aligning models to human reading times might promote a more uniform information distribution, though further research is needed to explore the connection between the UID hypothesis and alignment with reading times.

7 Discussion

We now discuss the broader implications of our results with respect to both cognitive modeling and natural language processing. One recurring theme in this paper is the relationship between $p_{\boldsymbol{\theta}}$, a probability distribution over strings estimated from large corpora, and p_{H}, the distribution implicated during cognitive tasks. In terms of cognitive modeling, it is widely accepted that it is useful to obtain a

[9] We use the LM Evaluation Harness (Gao et al., 2024).

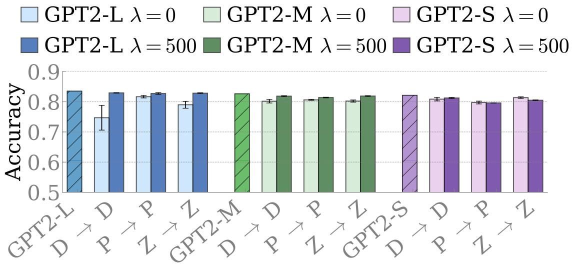


Figure 5: **Results for BLiMP.** Non-fine-tuned models are shown with hatching. Error bars are standard errors across random seeds. Fine-tuning leads to a decrease in accuracy. For results on all data splits, see Fig. 7.

Model	$\lambda = 0$		$\lambda = 500$	
	↓ UID$_v$	↑ 1-Gram%	↓ UID$_v$	↑ 1-Gram%
GPT2-L	6.69	84.84	6.69	84.84
D → D	$10.17_{2.93}$	$67.07_{14.68}$	$8.05_{0.77}$	$82.38_{0.18}$
P → P	$11.27_{1.74}$	$68.91_{1.76}$	$5.14_{0.35}$	$76.75_{0.79}$
Z → Z	$13.85_{0.98}$	$84.60_{1.91}$	$5.61_{0.09}$	$81.12_{0.20}$
GPT2-M	7.67	84.39	7.67	84.39
D → D	$8.94_{0.25}$	$80.82_{0.35}$	$6.63_{0.35}$	$83.22_{1.08}$
P → P	$7.81_{0.46}$	$67.51_{0.78}$	$4.54_{0.17}$	$75.91_{0.85}$
Z → Z	$12.12_{1.42}$	$84.95_{0.86}$	$5.88_{0.05}$	$81.94_{0.91}$
GPT2-S	5.70	82.53	5.70	82.53
D → D	$6.27_{0.52}$	$81.14_{0.95}$	$4.50_{0.11}$	$76.56_{2.02}$
P → P	$6.91_{0.25}$	$72.17_{1.99}$	$4.11_{0.10}$	$69.54_{0.43}$
Z → Z	$7.79_{1.02}$	$79.61_{1.94}$	$4.97_{0.18}$	$75.92_{0.68}$

Table 3: Surprisal variance (UID$_v$), and unique unigram ratios (1-Gram%) of model generated completions. See Tab. 11 for results on all data splits.

$p_{\boldsymbol{\theta}}$ that is as close as possible to p_{H} in order to test and refine psychological theories. While previous work has proposed alignment procedures between LMs and brain data (Toneva and Wehbe, 2019), within the realm of psycholinguistics, researchers have generally opted for whatever model is the current state of the art. Only recently have contributions started to search for which combinations of training data, model size, and model architecture produce the best cognitive alignment (Oh and Schuler, 2023a; Steuer et al., 2023). Our results suggest an alternative to this search by directly aligning language models to psychometric data. We are optimistic that such aligned models will enable a more precise evaluation of psycholinguistic theories. Of course, it remains an open question to what extent alignment on reading generalizes to other cognitive tasks. We have discussed p_{H} in terms of a single hypothesized construct, but it is likely that people's predictions change with task demands, e.g., when skimming versus proofreading a text. Testing whether our results hold for other types of psychometric predictive tasks is an important question for future research.

The main technical contribution of this work is a technique that aligns language models more closely to human psychometric data. Inspired by the implicit parameter optimization in DPO, our approach goes beyond the original Bradley–Terry assumption and demonstrates that it is possible to fine-tune under other implicit statistical models. Exploring a wider range of such models goes hand in hand with exploring new sources of human psychometric data, as we have done here with reading times. Psychologists have devised methods for collecting a diverse array of cognitive signals, including EEG, fMRI, mouse-tracking, and self-paced reading, to name a few. Aligning models, using the method proposed here, on such data or combinations thereof, will be an important next step in this research.

Finally, what do our results say about the relationship between cognitive modeling and other domains in NLP? At first glance, they seem to suggest that alignment with reading times is not an effective strategy to broadly boost performance on NLP tasks. Although KL regularization reduces the increase of the language model's perplexity on held-out data, we still observe decreased performance on downstream NLP tasks. However, given the lack of reading time data for LM-generated text, we relied on pre-existing eye-tracking datasets. Future studies could experiment with text generated by language models, particularly as new eye-tracking datasets for LM-generated text are being developed (Bolliger et al., 2024). Furthermore, Rafailov et al. (2023) derive the optimal solution to the KL regularized RL objective in DPO, while our study excluded the KL term when deriving optimal coefficients. Future work could investigate approaches closer to DPO to compute the optimal coefficients.

8 Conclusion

We have presented a novel technique to align language models with human reading data by implicitly optimizing the parameters of a linear regression model. Our experiments on held-out test sets show our method reliably improves the predictive power of language models with various parameter counts on human reading times. Furthermore, our findings confirm previous research on the inverse relationship between perplexity and psychometric predictive power. We believe our results pave the way for better assessment of psychological theories using more cognitively aligned language models.

Limitations

Our study has several limitations. First, we only include predictors for the current unit. Future research could explore the impact of adding previous units' predictors as well as additional predictors, such as contextual entropy, as suggested in Pimentel et al. (2023), to the objective. Second, our study only tests and evaluates English language data. Expanding these studies to a wider variety of languages is an important next step in establishing the generalizability of our methods. Third, we estimate surprisal based on individual sentences, whereas surprisal estimated for whole paragraphs may yield more accurate estimates due to the additional context.

Ethics Statement

Our study introduces a technique for aligning language models to human psychometric data. When working with human psychometric data, specifically eye-tracking data, it is important to consider potential privacy risks (Jäger et al., 2020; Lohr and Komogortsev, 2022). In this study, we used well-established datasets (Kennedy et al., 2003; Luke and Christianson, 2018; Hollenstein et al., 2018, 2020), where personal identifying information had been anonymized prior to our access. Additionally, we are aware of the potential biases inherent in language models and human reading data. Our goal is to ensure that our models and evaluations do not propagate or amplify existing biases.

Acknowledgments

We would like to thank the ARR reviewers and area chair for their valuable feedback and comments. Afra Amini is supported by ETH AI Center doctoral fellowship. David Robert Reich is supported by the Swiss National Science Foundation (SNSF) under grant IZCOZ0_220330/1 (EyeNLG).

References

David A. Balota, Alexander Pollatsek, and Keith Rayner. 1985. The interaction of contextual constraints and parafoveal visual information in reading. *Cognitive Psychology*, 17(3).

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

Lena Sophia Bolliger, Patrick Haller, Isabelle Caroline Rose Cretton, David Robert Reich, Tannon Kew, and Lena Ann Jäger. 2024. Emtec: A corpus of eye movements on machine-generated texts. *Preprint*, arXiv:2408.04289.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4).

Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30.

Thomas Hikaru Clark, Clara Meister, Tiago Pimentel, Michael Hahn, Ryan Cotterell, Richard Futrell, and Roger Levy. 2023. A cross-linguistic pressure for Uniform Information Density in word order. *Transactions of the Association for Computational Linguistics*, 11.

Katherine A. DeLong, Thomas P. Urbach, and Marta Kutas. 2005. Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, 8.

Shuwen Deng, Paul Prasse, David Reich, Tobias Scheffer, and Lena Jäger. 2023. Pre-trained language models augmented with synthetic scanpaths for natural language understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore. Association for Computational Linguistics.

B. Efron. 1979. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1).

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

August Fenk and Gertraud Fenk. 1980. Konstanz im Kurzzeitgedächtnis - Konstanz im sprachlichen Informationsfluß? *Zeitschrift für experimentelle und angewandte Psychologie*, 27(3).

Ira Fischler and Paul A. Bloom. 1979. Automatic and attentional processes in the effects of sentence contexts on word recognition. *Journal of Verbal Learning and Verbal Behavior*, 18(1).

Stefan L. Frank and Rens Bod. 2011. Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, 22(6).

Lyn Frazier and Keith Rayner. 1982. Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, 14(2).

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. A framework for few-shot language model evaluation.

Mario Giulianelli, Luca Malagutti, Juan Luis Gastaldi, Brian DuSell, Tim Vieira, and Ryan Cotterell. 2024. On the proper treatment of tokenization in psycholinguistics. *Preprint*, arXiv:2410.02691.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics*, Salt Lake City, Utah. Association for Computational Linguistics.

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*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28.

Arthur E. Hoerl and Robert W. Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1).

Nora Hollenstein, Federico Pirovano, Ce Zhang, Lena Jäger, and Lisa Beinborn. 2021. Multilingual language models predict human reading behavior. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online. Association for Computational Linguistics.

Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. 2018. ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Scientific Data*, 5.

Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. 2020. ZuCo 2.0: A dataset of physiological recordings during natural reading and annotation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association.

Jacob Louis Hoover, Morgan Sonderegger, Steven T. Piantadosi, and Timothy J. O'Donnell. 2023. The Plausibility of Sampling as an Algorithmic Theory of Sentence Processing. *Open Mind*, 7.

Lena A. Jäger, Silvia Makowski, Paul Prasse, Sascha Liehr, Maximilian Seidler, and Tobias Scheffer. 2020. Deep Eyedentification: biometric identification using micro-movements of the eye. In *Machine Learning and Knowledge Discovery in Databases*, Cham. Springer International Publishing.

Marcel A. Just and Patricia A. Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4).

Alan Kennedy, Robin Hill, and Joël Pynte. 2003. The Dundee corpus. In *In Proceedings of the 12th European Conference on Eye Movement*.

Gina R. Kuperberg and T. Florian Jaeger. 2016. What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, 31(1).

Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, and Kentaro Inui. 2021. Lower perplexity is not always human-like. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online. Association for Computational Linguistics.

Marta Kutas and Steven A. Hillyard. 1984. Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307.

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

Roger Levy and T. Florian Jaeger. 2006. Speakers optimize information density through syntactic reduction. In *Advances in Neural Information Processing Systems*, volume 19, Cambridge, MA, USA. MIT Press.

Xiaochen Li, Zheng-Xin Yong, and Stephen H. Bach. 2024. Preference tuning for toxicity mitigation generalizes across languages. *Preprint*, arXiv:2406.16235.

Dillon Lohr and Oleg V. Komogortsev. 2022. Eye know you too: Toward viable end-to-end eye movement biometrics for user authentication. *IEEE Transactions on Information Forensics and Security*, 17.

Ilya Loshchilov and Frank Hutter. 2017. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*.

Steven G. Luke and Kiel Christianson. 2018. The provo corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 50.

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the uniform information density hypothesis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online. Association for Computational Linguistics.

George A. Miller and Kathryn Ojemann McKean. 1964. A chronometric study of some relations between sentences. *Quarterly Journal of Experimental Psychology*, 16(4).

Byung-Doh Oh, Christian Clark, and William Schuler. 2022. Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence*, 5.

Byung-Doh Oh and William Schuler. 2023a. 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*, Singapore. Association for Computational Linguistics.

Byung-Doh Oh and William Schuler. 2023b. 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.

Byung-Doh Oh and William Schuler. 2024. Leading whitespaces of language models' subword vocabulary poses a confound for calculating word probabilities. *Preprint*, arXiv:2406.10851.

Andreas Opedal, Eleanor Chodroff, Ryan Cotterell, and Ethan Gotlieb Wilcox. 2024. On the role of context in reading time prediction. *Preprint*, arXiv:2409.08160.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany. Association for Computational Linguistics.

Tiago Pimentel and Clara Meister. 2024. How to compute the probability of a word. *Preprint*, arXiv:2406.14561.

Tiago Pimentel, Clara Meister, Ethan G. Wilcox, Roger P. Levy, and Ryan Cotterell. 2023. On the Effect of Anticipation on Reading Times. *Transactions of the Association for Computational Linguistics*, 11.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8).

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3).

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *Preprint*, arXiv:1707.06347.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada. Association for Computational Linguistics.

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

Cory Shain and William Schuler. 2021. Continuous-time deconvolutional regression for psycholinguistic modeling. *Cognition*, 215.

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

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

Julius Steuer, Marius Mosbach, and Dietrich Klakow. 2023. Large GPT-like models are bad babies: A closer look at the relationship between linguistic competence and psycholinguistic measures. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, Singapore. Association for Computational Linguistics.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33.

Mariya Toneva and Leila Wehbe. 2019. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Neural Information Processing Systems*.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8.

Ethan Wilcox, Clara Meister, Ryan Cotterell, and Tiago Pimentel. 2023a. Language model quality correlates with psychometric predictive power in multiple languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore. Association for Computational Linguistics.

Ethan G. Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023b. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the Cognitive Science Society*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online. Association for Computational Linguistics.

Duo Yang and Nora Hollenstein. 2023. Plm-as: Pretrained language models augmented with scanpaths for sentiment classification. *Proceedings of the Northern Lights Deep Learning Workshop*.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-tuning language models from human preferences. *Preprint*, arXiv:1909.08593.

A Derivations

A.1 Deriving Optimal Coefficients

We now discuss how we estimate the optimal coefficient $\boldsymbol{\beta}^\star$ to approximate the reward, as described in §3.2. We start with finding the optimal coefficients

$$\boldsymbol{\beta}_{\boldsymbol{\theta}}^\star = \underset{\boldsymbol{\beta} \in \mathbb{R}^D}{\operatorname{argmin}} \frac{1}{N} \|\boldsymbol{\psi} - \mathbf{X}_{\boldsymbol{\theta}} \boldsymbol{\beta}_{\boldsymbol{\theta}}\|^2. \tag{14}$$

Assuming $\left(\mathbf{X}_{\boldsymbol{\theta}}^\top \mathbf{X}_{\boldsymbol{\theta}}\right)^{-1}$ is invertible, we take the derivative with respect to $\boldsymbol{\beta}_{\boldsymbol{\theta}}$, set it to 0 and solve for $\boldsymbol{\beta}_{\boldsymbol{\theta}}^\star$:

$$\begin{aligned}
& -\frac{2}{N} \mathbf{X}_{\boldsymbol{\theta}}^\top \left(\boldsymbol{\psi} - \mathbf{X}_{\boldsymbol{\theta}} \boldsymbol{\beta}_{\boldsymbol{\theta}}^\star\right) &= \quad & 0 && (15a)\\
\Leftrightarrow \quad & -\mathbf{X}_{\boldsymbol{\theta}}^\top \boldsymbol{\psi} + \mathbf{X}_{\boldsymbol{\theta}}^\top \mathbf{X}_{\boldsymbol{\theta}} \boldsymbol{\beta}_{\boldsymbol{\theta}}^\star &= \quad & 0 && (15b)\\
\Leftrightarrow \quad & \mathbf{X}_{\boldsymbol{\theta}}^\top \mathbf{X}_{\boldsymbol{\theta}} \boldsymbol{\beta}_{\boldsymbol{\theta}}^\star &= \quad & \mathbf{X}_{\boldsymbol{\theta}}^\top \boldsymbol{\psi} && (15c)\\
\Leftrightarrow \quad & \boldsymbol{\beta}_{\boldsymbol{\theta}}^\star &= \quad & \left(\mathbf{X}_{\boldsymbol{\theta}}^\top \mathbf{X}_{\boldsymbol{\theta}}\right)^{-1} \mathbf{X}_{\boldsymbol{\theta}}^\top \boldsymbol{\psi}. && (15d)
\end{aligned}$$

In theory, $\left(\mathbf{X}_{\boldsymbol{\theta}}^\top \mathbf{X}_{\boldsymbol{\theta}}\right)^{-1}$ may not always be invertible, which is why we add a regularization term $\rho\mathbf{I}$, where $\mathbf{I}$ is the identity matrix and $\rho > 0$ is a parameter determining the strength of the regularization. The resulting estimator $\boldsymbol{\beta}_{\boldsymbol{\theta}}^\star = \left(\mathbf{X}_{\boldsymbol{\theta}}^\top \mathbf{X}_{\boldsymbol{\theta}} + \rho\mathbf{I}\right)^{-1} \mathbf{X}_{\boldsymbol{\theta}}^\top \boldsymbol{\psi}$ is known as the ridge regression estimator (Hoerl and Kennard, 1970) and presents the solution to the following problem:

$$\boldsymbol{\beta}_{\boldsymbol{\theta}}^\star = \underset{\boldsymbol{\beta}_{\boldsymbol{\theta}} \in \mathbb{R}^D}{\operatorname{argmin}} \frac{1}{N} \|\boldsymbol{\psi} - \mathbf{X}_{\boldsymbol{\theta}} \boldsymbol{\beta}_{\boldsymbol{\theta}}\|^2 + \gamma \|\boldsymbol{\beta}_{\boldsymbol{\theta}}\|^2, \tag{16}$$

where $\gamma > 0$. We define $\rho = N\gamma$. Then setting the derivative of Eq. (16) to zero leads to

$$\begin{aligned}
& -\frac{2}{N} \mathbf{X}_{\boldsymbol{\theta}}^\top \left(\boldsymbol{\psi} - \mathbf{X}_{\boldsymbol{\theta}} \boldsymbol{\beta}_{\boldsymbol{\theta}}^\star\right) + 2\gamma \boldsymbol{\beta}_{\boldsymbol{\theta}}^\star &= \quad & 0 && (17a)\\
\Leftrightarrow \quad & \frac{2}{N} \mathbf{X}_{\boldsymbol{\theta}}^\top \mathbf{X}_{\boldsymbol{\theta}} \boldsymbol{\beta}_{\boldsymbol{\theta}}^\star + 2\gamma \boldsymbol{\beta}_{\boldsymbol{\theta}}^\star &= \quad & \frac{2}{N} \mathbf{X}_{\boldsymbol{\theta}}^\top \boldsymbol{\psi} && (17b)\\
\Leftrightarrow \quad & \mathbf{X}_{\boldsymbol{\theta}}^\top \mathbf{X}_{\boldsymbol{\theta}} \boldsymbol{\beta}_{\boldsymbol{\theta}}^\star + N\gamma \boldsymbol{\beta}_{\boldsymbol{\theta}}^\star &= \quad & \mathbf{X}_{\boldsymbol{\theta}}^\top \boldsymbol{\psi} && (17c)\\
\Leftrightarrow \quad & \mathbf{X}_{\boldsymbol{\theta}}^\top \mathbf{X}_{\boldsymbol{\theta}} \boldsymbol{\beta}_{\boldsymbol{\theta}}^\star + \rho \boldsymbol{\beta}_{\boldsymbol{\theta}}^\star &= \quad & \mathbf{X}_{\boldsymbol{\theta}}^\top \boldsymbol{\psi} && (17d)\\
\Leftrightarrow \quad & \boldsymbol{\beta}_{\boldsymbol{\theta}}^\star &= \quad & \left(\mathbf{X}_{\boldsymbol{\theta}}^\top \mathbf{X}_{\boldsymbol{\theta}} + \rho\mathbf{I}\right)^{-1} \mathbf{X}_{\boldsymbol{\theta}}^\top \boldsymbol{\psi}. && (17e)
\end{aligned}$$

A.2 Solving for Optimal Coefficients

To compute the regression coefficients efficiently, we use the Cholesky decomposition to solve for $\boldsymbol{\beta}_{\boldsymbol{\theta}}^\star$ given as

$$\boldsymbol{\beta}_{\boldsymbol{\theta}}^\star = (\mathbf{X}_{\boldsymbol{\theta}}^\top \mathbf{X}_{\boldsymbol{\theta}} + \rho\mathbf{I})^{-1} \mathbf{X}_{\boldsymbol{\theta}}^\top \boldsymbol{\psi}, \tag{18}$$

where ρ is the regularization parameter, which we set $\rho = 1\mathrm{e} - 5$ and $\mathbf{I}$ is the identity matrix. Since $\mathbf{X}_{\boldsymbol{\theta}}^\top \mathbf{X}_{\boldsymbol{\theta}} + \rho\mathbf{I}$ is symmetric and positive definite, we compute the Cholesky decomposition

$$\mathbf{X}_{\boldsymbol{\theta}}^\top \mathbf{X}_{\boldsymbol{\theta}} + \rho\mathbf{I} = \mathbf{L}\mathbf{L}^\top, \tag{19}$$

where $\mathbf{L} \in \mathbb{R}^{D \times D}$ is a lower triangular matrix. To solve for $\boldsymbol{\beta}_{\boldsymbol{\theta}}^\star$, we first solve for an intermediate vector $\mathbf{z}$

$$\mathbf{L}\mathbf{z} = \mathbf{X}_{\boldsymbol{\theta}}^\top \boldsymbol{\psi} \tag{20}$$

We then solve for $\boldsymbol{\beta}_{\boldsymbol{\theta}}^\star$

$$\mathbf{L}^\top \boldsymbol{\beta}_{\boldsymbol{\theta}}^\star = \mathbf{z}. \tag{21}$$

B Datasets, Reading Times & Parameters

B.1 Datasets

Here, we provide additional details on the datasets and reading time measurements used during our analysis. We fine-tune and evaluate models on the Dundee (Kennedy et al., 2003), Provo (Luke and Christianson, 2018), and ZuCo corpora. For the ZuCo corpus, we use data from tasks 1 and 2 from the ZuCo 1.0 corpus (Hollenstein et al., 2018) and task 1 from the ZuCo 2.0 corpus (Hollenstein et al., 2020). All data used in our analysis is publicly available. For the Dundee and ZuCo corpora, we process the data used by Hollenstein et al. (2021), which contains word-level means for fixation counts and reading durations, averaged over all participants, and split into individual sentences.[10] For the Provo corpus, we compute the mean reading times from the official repository. [11] From all datasets, we remove duplicate sentences and short sentences with less than four words. The mean number of sentences and words for train and test sets are given in Tab. 4.

Configuration	Train Tokens	Train Sents	Test Tokens	Test Sents
Dundee (D) → Provo (P)	20894.7	980	1144	54
Dundee (D) → Dundee (D)	20894.7	980	20207	931
Dundee (D) → ZuCo (Z)	20894.7	980	7715	424
ZuCo (Z) → Provo (P)	7761	451	1144	54
ZuCo (Z) → Dundee (D)	7761	451	20207	931
ZuCo (Z) → ZuCo (Z)	7761	451	7715	424
Provo (P) → Provo (P)	1113.7	56	1144	54
Provo (P) → Dundee (D)	1113.7	56	20207	931
Provo (P) → ZuCo (Z)	1113.7	56	7715	424

Table 4: Data splits and configurations for fine-tuning and evaluation. Numbers indicate the **mean** number of tokens and sentences in each train and test split across random seeds.

B.2 Reading Times

For Dundee and ZuCo, we extract the mean first pass duration over the participants, which is defined as "the sum of all fixations on w from the first time a subject fixates w to the first time the subject fixates another token"(Hollenstein et al., 2021, p.109). Similarly, for Provo, we compute the mean gaze duration defined as the "Dwell time (i.e., summation of the duration across all fixations) of the first run within the current interest area" (Luke and Christianson, 2018, Tab. 2). While these two definitions are very similar, they may not be exactly identical. In Tab. 5, we compare the mean and standard deviation of reading times as well as the number of zero reading times. We observe that while Dundee and Provo exhibit relatively similar means and standard deviations, ZuCo shows overall lower mean reading times. Unlike Dundee and ZuCo, Provo contains no instances of words with zero reading times, likely due to the high number of participants.

Dataset	Mean Reading Times	STD Reading Times	Zero Count
Dundee	140.59	88.49	854
Provo	164.16	77.77	0
ZuCo	92.28	52.21	176

Table 5: Mean, standard deviation, and zeros count for the reading times from Dundee, Provo, and ZuCo.

B.3 Fine-Tuning Parameters

For all runs, we use the parameter configurations in Tab. 6 and repeat each experiment 3 times with random seeds (42, 8, and 64). We fine-tune and evaluate three GPT-2 models: GPT-2 Small with 117 million parameters, GPT-2 Medium with 345 million parameters, and GPT-2 Large with 762 million

[10] https://github.com/DS3Lab/multilingual-gaze

[11] https://osf.io/sjefs/

parameters. All models are fine-tuned for 5k data steps using a batch size of 1 and gradient accumulation of 2, leading to a total of 2.5k optimization steps. To adjust the learning rate during fine-tuning, we use a cosine annealing learning rate schedule[12] (Loshchilov and Hutter, 2017) with a maximum learning rate of 1.5e-5, minimum learning rate of 2e-7. During each cycle, we decrease the learning by a factor of 0.8 and increase the cycle length by a factor of 1.8. For all experiments, we use NVIDIA GeForce RTX 3090, RTX 4090, and RTX 2080 Ti GPUs. The GPU times vary depending on the model's parameter counts and the size of the data splits, taking a minimum of approximately 12 minutes for GPT-2 Small fine-tuned on Provo and evaluated on Provo and a maximum of 4 hours for GPT-2 Large fine-tuned on Dundee and evaluated on Dundee.

Parameter	Setting
Optimizer	AdamW
Scheduler	Cosine Annealing With Warm Restarts
Batch Size	1
Grad. Accumulation	2
Total Steps	5000
Optimizer Steps	2500
Max Learning Rate	1.5e-5
Min Learning Rate	2.0e-7
Decrease Rate of Max Learning Rate	0.8
Cycle Steps Magnification	1.8
Warm Up Steps	100

Table 6: Fine-tuning parameters used consistently across different runs.

B.4 Evaluation

We perform evaluation every 50 steps on the held-out test splits in Tab. 4, during which we compute the surprisal estimates, the regressors' coefficients, and the language model's perplexity for each batch. Using the surprisal estimates we compute the Δ_{llh} by performing a 5-fold cross-validation on the test data, where we fit baseline and target linear regressors using ordinary least squares.[13] Note that we do not scale the predictor variables to maintain consistency between the linear regression and the calculation of our reward in Eq. (13), where batch-level surprisal estimates prevent global scaling.

[12] https://github.com/katsura-jp/pytorch-cosine-annealing-with-warmup

[13] https://www.statsmodels.org/stable/generated/statsmodels.regression.linear_model.OLS.html

C Detailed Results

C.1 Δ_{llh} Change

Data	GPT2-L $\Delta_{\text{llh}}^{\text{start}}$	GPT2-L $\Delta_{\text{llh}}^{\text{max}}$	GPT2-L % Increase	GPT2-M $\Delta_{\text{llh}}^{\text{start}}$	GPT2-M $\Delta_{\text{llh}}^{\text{max}}$	GPT2-M % Increase	GPT2-S $\Delta_{\text{llh}}^{\text{start}}$	GPT2-S $\Delta_{\text{llh}}^{\text{max}}$	GPT2-S % Increase
D → D	0.63	0.98 ± 0.09	55.29	0.68	0.90 ± 0.06	32.82	0.78	1.11 ± 0.01	42.88
P → D	0.63	0.78 ± 0.04	23.29	0.68	0.84 ± 0.00	24.23	0.78	1.01 ± 0.02	29.90
Z → D	0.63	0.69 ± 0.01	9.08	0.68	0.75 ± 0.03	11.08	0.78	0.91 ± 0.04	16.60
D → P	1.67	2.19 ± 0.37	31.36	2.04	2.14 ± 0.05	4.93	2.14	2.52 ± 0.09	17.78
P → P	1.67	2.43 ± 0.14	45.27	2.04	2.36 ± 0.10	15.62	2.14	2.67 ± 0.15	24.77
Z → P	1.67	1.71 ± 0.03	2.12	2.04	2.18 ± 0.03	6.46	2.14	2.18 ± 0.01	2.00
D → Z	1.17	1.59 ± 0.15	36.58	1.37	1.53 ± 0.06	11.95	1.41	1.75 ± 0.06	24.50
P → Z	1.17	1.34 ± 0.09	15.01	1.37	1.45 ± 0.08	5.73	1.41	1.69 ± 0.07	20.09
Z → Z	1.17	2.12 ± 0.10	81.39	1.37	1.94 ± 0.04	41.80	1.41	2.26 ± 0.14	61.20

Table 7: Mean start and maximum $\Delta_{\text{llh}}(10^{-2}$ nats) for Tab. 2, including standard errors across random seeds, rounded to two decimal places.

C.2 Perplexity and Δ_{llh}

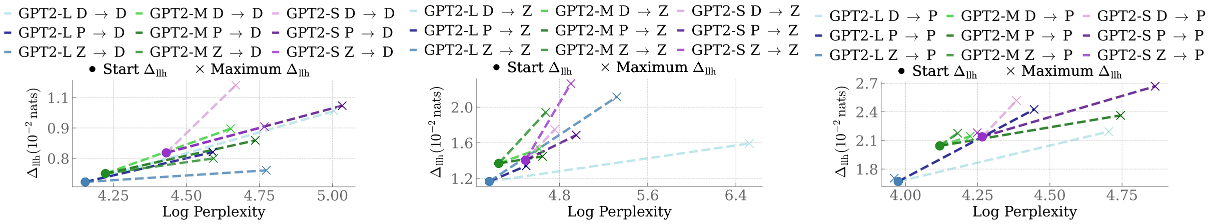


Figure 6: Relationship between log perplexity and $\Delta_{\text{llh}}(10^{-2}$ nats) for all models and data splits. Increases in Δ_{llh} correspond to increases in perplexity.

C.3 BLiMP

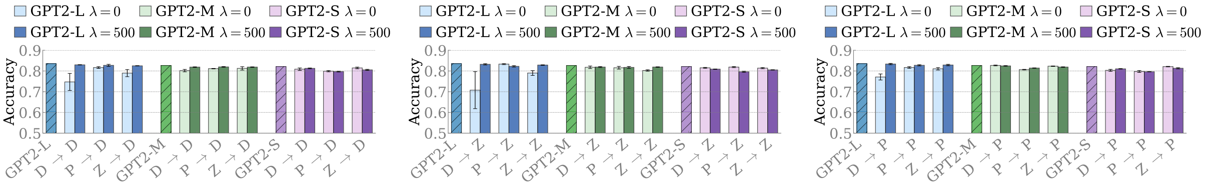


Figure 7: Results for all models on BLiMP. Baseline (i.e., non-fine-tuned) models are shown with hatching. Error bars are standard errors across the three random seeds. We observe a drop in accuracy as a result of fine-tuning.

C.4 Narrative Understanding

To measure models' abilities to track entities and produce text that is consistent with narrative structure, we evaluate them on LAMBADA (Paperno et al., 2016). This dataset requires that models produce the final word of a narrative. People can easily achieve higher performance on this task if they are given the full narrative context, but not if they are only given the previous sentence. Thus, performing well requires an understanding of broader contexts. The performance of our models is visualized in Fig. 8. As with BLiMP, we find that fine-tuned models perform slightly worse at this task compared to non-fine-tuned baselines.

C.5 Results for the Regularized Objective

In this section, we present additional results for fine-tuning models using the KL regularized objective with a regularization coefficient of $\lambda = 500$. As shown in Tab. 8, the maximum Δ_{llh} values and percentage

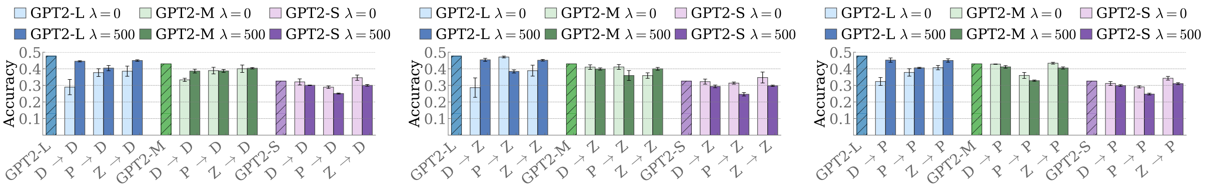


Figure 8: **Results for LAMBADA.** Baseline (i.e., non-fine-tuned) models are shown with hatching. Error bars are standard errors across the three random seeds. As with BLiMP, performance tends to decrease with fine-tuning.

increases tend to be lower compared to those for the unregularized objective in Tab. 7. However, standard errors for the maximum Δ_{llh} are consistently lower, suggesting that fine-tuning is more stable across random seeds when regularizing the reward. Additionally, when visualizing the MSE and Δ_{llh} throughout fine-tuning (Fig. 9), we observe more consistent improvements for Dundee → Provo and ZuCo → Provo compared to the trajectories for the unregularized objective in Fig. 1.

	GPT2-L			GPT2-M			GPT2-S		
Data	$\Delta_{\text{llh}}^{\text{start}}$	$\Delta_{\text{llh}}^{\text{max}}$	% Increase	$\Delta_{\text{llh}}^{\text{start}}$	$\Delta_{\text{llh}}^{\text{max}}$	% Increase	$\Delta_{\text{llh}}^{\text{start}}$	$\Delta_{\text{llh}}^{\text{max}}$	% Increase
D → D	0.63	0.81 ± 0.00	28.27	0.68	0.85 ± 0.03	25.49	0.78	1.02 ± 0.01	30.72
P → D	0.63	0.71 ± 0.01	12.69	0.68	0.71 ± 0.00	5.62	0.78	0.93 ± 0.01	19.34
Z → D	0.63	0.69 ± 0.00	8.63	0.68	0.71 ± 0.02	5.24	0.78	0.88 ± 0.00	13.22
D → P	1.67	2.09 ± 0.03	24.91	2.04	2.18 ± 0.03	6.87	2.14	2.36 ± 0.03	10.47
P → P	1.67	1.98 ± 0.04	18.70	2.04	2.25 ± 0.07	10.00	2.14	2.37 ± 0.02	10.95
Z → P	1.67	1.86 ± 0.03	11.13	2.04	2.19 ± 0.00	7.38	2.14	2.19 ± 0.01	2.58
D → Z	1.17	1.36 ± 0.03	16.86	1.37	1.50 ± 0.07	9.52	1.40	1.69 ± 0.02	20.01
P → Z	1.17	1.35 ± 0.04	16.14	1.37	1.52 ± 0.01	11.24	1.40	1.76 ± 0.03	25.34
Z → Z	1.17	1.58 ± 0.07	35.24	1.37	1.78 ± 0.01	29.88	1.40	1.98 ± 0.06	41.12

Table 8: Mean start and maximum $\Delta_{\text{llh}}(10^{-2}$ nats) values using the **KL regularized objective** with $\lambda = 500$, including standard errors across random seeds, rounded to two decimal places.

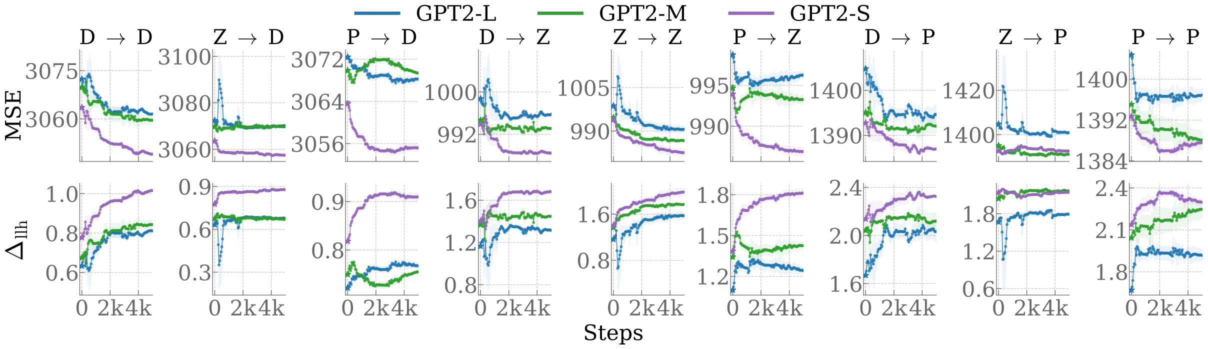


Figure 9: MSE and $\Delta_{\text{llh}}(10^{-2}$ nats) changes for all data splits using the **KL regularized objective** with $\lambda = 500$. Bands show the standard error across seeds. Compared to the results for the unregularized objective in §5.1, we observe smaller but consistent decreases in MSE and increases in Δ_{llh} for almost all configurations.

D Reading Times

D.1 Total Reading Duration & First Fixation Duration

Throughout this work, we have focused on predicting gaze durations. Here, we extend our analysis by fine-tuning models to predict total reading durations—which are the summed durations of all fixations on a unit u—and first fixation durations, which are the durations of the first fixation on a unit u. As shown in Tab. 9, models predicting total reading durations start at higher Δ_{llh} values, while models predicting first fixation durations start with lower Δ_{llh} values compared to those predicting gaze durations. On average, we observe lower percentage increases for total reading durations and lower increases for first fixation durations. Similar to the trajectories for gaze duration in Fig. 1, the trajectories for total reading durations in Fig. 10 show decreasing MSE and increasing Δ_{llh} for most configurations. The trajectories for the first fixation durations (Fig. 11) are less consistent, particularly for models fine-tuned on the ZuCo corpus, where the MSE tends to increase throughout fine-tuning.

	GPT2-L			GPT2-M			GPT2-S		
Data	$\Delta_{\text{llh}}^{\text{start}}$	$\Delta_{\text{llh}}^{\text{max}}$	% Increase	$\Delta_{\text{llh}}^{\text{start}}$	$\Delta_{\text{llh}}^{\text{max}}$	% Increase	$\Delta_{\text{llh}}^{\text{start}}$	$\Delta_{\text{llh}}^{\text{max}}$	% Increase
Total Reading Duration									
D → D	1.11	1.93 ± 0.33	73.20	1.18	1.62 ± 0.07	37.48	1.35	1.83 ± 0.07	35.41
P → D	1.11	1.22 ± 0.04	9.20	1.18	1.38 ± 0.03	17.26	1.35	1.60 ± 0.01	18.62
Z → D	1.11	1.28 ± 0.07	15.01	1.18	1.27 ± 0.01	7.83	1.35	1.54 ± 0.04	14.47
D → P	3.11	3.79 ± 0.20	21.90	3.45	3.63 ± 0.06	5.29	3.86	4.27 ± 0.05	10.61
P → P	3.11	4.16 ± 0.16	33.64	3.45	4.29 ± 0.11	24.43	3.86	4.59 ± 0.16	18.70
Z → P	3.11	3.89 ± 0.33	24.95	3.45	3.50 ± 0.02	1.42	3.86	3.91 ± 0.00	1.33
D → Z	2.82	3.38 ± 0.54	19.93	2.98	2.99 ± 0.02	0.19	2.89	3.25 ± 0.19	12.47
P → Z	2.82	2.85 ± 0.01	1.31	2.98	3.03 ± 0.01	1.64	2.89	2.98 ± 0.01	3.10
Z → Z	2.82	5.28 ± 0.79	87.40	2.98	3.92 ± 0.14	31.35	2.89	4.23 ± 0.23	46.37
First Fixation Duration									
D → D	0.08	1.21 ± 0.47	1397.26	0.07	0.12 ± 0.02	64.10	0.09	0.14 ± 0.01	52.02
P → D	0.08	0.11 ± 0.02	35.25	0.07	0.13 ± 0.01	81.56	0.09	0.20 ± 0.04	121.22
Z → D	0.08	0.10 ± 0.02	25.05	0.07	0.07 ± 0.00	1.10	0.09	0.10 ± 0.01	9.49
D → P	1.29	5.07 ± 1.77	292.36	1.51	1.57 ± 0.13	3.87	1.37	1.55 ± 0.20	13.46
P → P	1.29	2.13 ± 0.08	64.61	1.51	2.50 ± 0.05	66.19	1.37	2.55 ± 0.31	86.55
Z → P	1.29	1.44 ± 0.41	11.26	1.51	1.51 ± 0.01	0.33	1.37	1.38 ± 0.02	1.01
D → Z	0.20	1.37 ± 0.46	591.38	0.21	0.31 ± 0.01	51.38	0.18	0.25 ± 0.02	40.97
P → Z	0.20	0.21 ± 0.01	8.20	0.21	0.23 ± 0.01	13.14	0.18	0.32 ± 0.06	75.49
Z → Z	0.20	0.55 ± 0.37	175.48	0.21	0.25 ± 0.08	21.36	0.18	0.28 ± 0.04	52.88

Table 9: Mean start and maximum $\Delta_{\text{llh}}(10^{-2}$ nats) values using **total reading durations and first fixation durations**, including standard errors across random seeds, rounded to two decimal places.

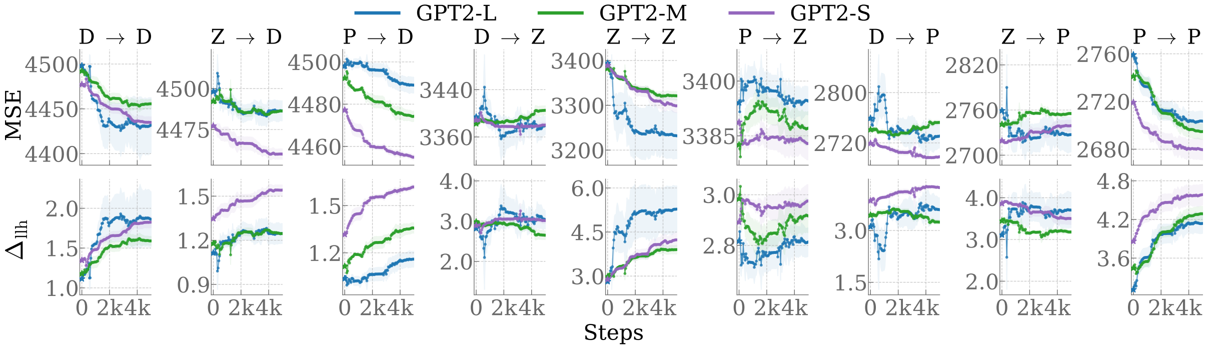


Figure 10: MSE and Δ_{llh} changes for **total reading durations**. Bands show standard errors across seeds.

D.2 Random Reading Times

Here we conduct additional experiments using random reading times to verify that the observed decreases in MSE and increases in Δ_{llh} are due to aligning language models with human reading times and not

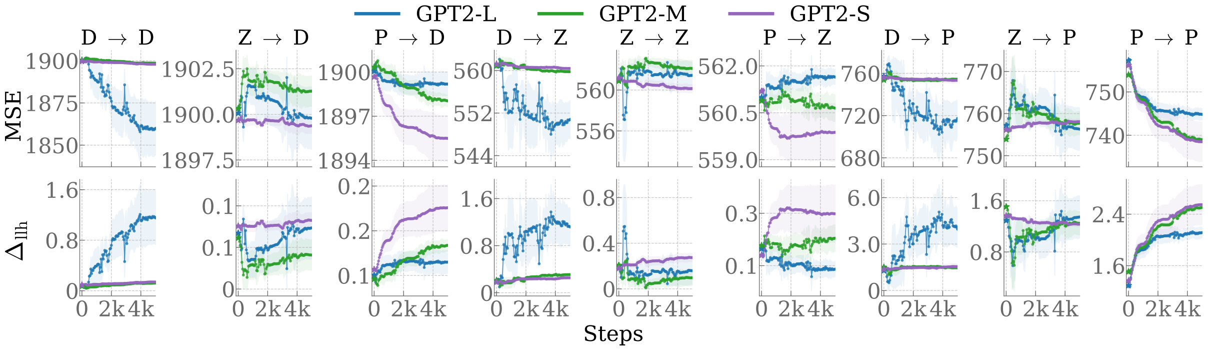


Figure 11: MSE and Δ_{llh} changes for **first fixation durations**. Bands show standard errors across seeds.

due to random noise or other confounding factors. Instead of fitting coefficients based on the reading times from the respective training dataset $\mathcal{D}$, we sample reading times from a Gaussian distribution, where the mean and standard deviation match those of the training dataset $\mathcal{D}$. The results in Fig. 12 show that fine-tuning on random reading times tends to have the opposite effect compared to fine-tuning on real reading times, leading to increasing MSE and decreasing Δ_{llh} throughout fine-tuning. Additionally, as in Tab. 7, we compare the start and maximum Δ_{llh} values, as well as the percentage increases. Tab. 10 shows that fine-tuning with random reading times only leads to minimal or no increases. These results show that random reading times do not improve language models at predicting reading times and confirm the effectiveness of our technique for aligning models to human reading times.

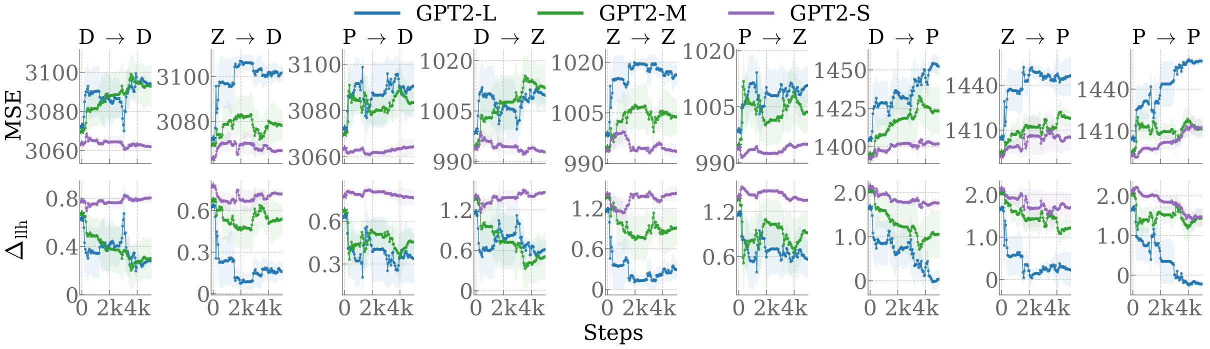


Figure 12: MSE and $\Delta_{\text{llh}}(10^{-2}$ nats) changes using **random reading times** sampled according to a normal distribution. Bands show the standard error across seeds. MSE increases, and Δ_{llh} decreases for most configurations, indicating that models become worse at predicting reading times throughout fine-tuning.

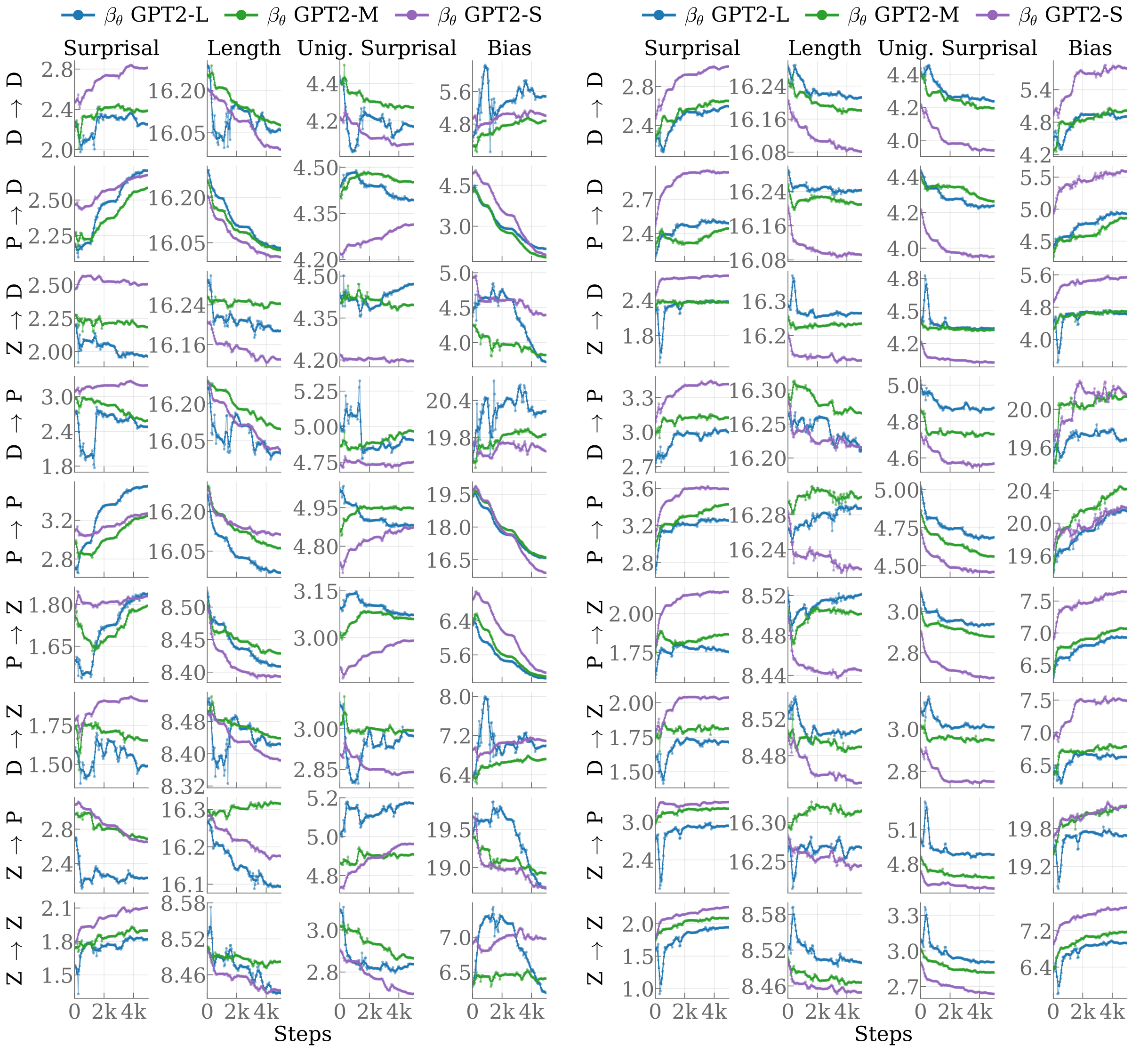
		GPT2-L			GPT2-M			GPT2-S	
Data	$\Delta_{\text{llh}}^{\text{start}}$	$\Delta_{\text{llh}}^{\text{max}}$	% Increase	$\Delta_{\text{llh}}^{\text{start}}$	$\Delta_{\text{llh}}^{\text{max}}$	% Increase	$\Delta_{\text{llh}}^{\text{start}}$	$\Delta_{\text{llh}}^{\text{max}}$	% Increase
D → D	0.63	0.67 ± 0.23	6.31	0.68	0.68 ± 0.00	0.00	0.78	0.83 ± 0.03	6.42
P → D	0.63	0.66 ± 0.00	3.84	0.68	0.68 ± 0.00	0.00	0.78	0.82 ± 0.03	5.67
Z → D	0.63	0.64 ± 0.01	0.88	0.68	0.68 ± 0.01	1.30	0.78	0.78 ± 0.00	0.00
D → P	1.67	1.67 ± 0.00	0.00	2.04	2.06 ± 0.01	0.56	2.14	2.14 ± 0.00	0.00
P → P	1.67	1.74 ± 0.01	4.06	2.04	2.04 ± 0.00	0.00	2.14	2.22 ± 0.02	3.87
Z → P	1.67	1.71 ± 0.02	2.44	2.04	2.08 ± 0.01	1.66	2.14	2.14 ± 0.01	0.03
D → Z	1.17	1.17 ± 0.00	0.00	1.37	1.37 ± 0.00	0.17	1.41	1.50 ± 0.04	6.48
P → Z	1.17	1.20 ± 0.02	3.20	1.37	1.37 ± 0.00	0.00	1.41	1.53 ± 0.01	9.14
Z → Z	1.17	1.20 ± 0.01	2.98	1.37	1.42 ± 0.03	3.40	1.41	1.45 ± 0.04	2.86

Table 10: Mean start and maximum $\Delta_{\text{llh}}(10^{-2}$ nats) values for **random reading times**, including standard errors across random seeds, rounded to two decimal places.

E Coefficients

Fig. 13a shows the trajectory of coefficients for all data splits during fine-tuning using the unregularized objective. While the results for unigram surprisal and bias coefficients are mixed, the coefficients for surprisal tend to increase during fine-tuning, and those for length tend to decrease. These trends are more consistent in experiments with the largest increases in Δ_{llh}, i.e., where the training and test splits come from the same datasets. In contrast, the coefficients in experiments involving domain shift (where the training and test splits come from different datasets) do not have such a clear pattern, particularly those without consistent increases in Δ_{llh} (Dundee → Provo and ZuCo → Provo).

Additionally, we visualize all coefficients from models trained with the regularized objective ($\lambda = 500$). As shown in Fig. 9, regularization leads to lower Δ_{llh} increases but more consistent trajectories across configurations, which is why we also hypothesize more consistent trajectories of the coefficients. As shown in Fig. 13b, the coefficients for surprisal and bias show a clear upward trend throughout fine-tuning, while the coefficients for unigram surprisal show a clear downward trend.



(a) Unregularized objective ($\lambda = 0$)

(b) Regularized objective ($\lambda = 500$)

Figure 13: **Mean coefficients of unit-level features over fine-tuning for all data splits.** Smoothed values (window size 5) are shown, with unsmoothed values in a pale version of the color.

F Text-Generation

Here we expand on the text-generation experiments described in §6.2. We sample 500 prefixes $\boldsymbol{c}$, each consisting of the first three words from the CNN/DailyMail dataset (Hermann et al., 2015; See et al., 2017), and generate completions $\boldsymbol{u} \sim p_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{c})$ of up to 50 tokens using our fine-tuned language models. For a completion $\boldsymbol{u}$ of length N, we write u_n to denote the n^{th} unit of $\boldsymbol{u}$. Further, let $\boldsymbol{c}_n$ be the context of u_n in $\boldsymbol{u}$, including the prefix $\boldsymbol{c}$. We exclude short completions with $|\boldsymbol{u}| < 3$ and estimate surprisal with a separate language model, Pythia-70m (Biderman et al., 2023) using the code from Pimentel and Meister (2024).[14] Recently, Oh and Schuler (2024); Pimentel and Meister (2024) have argued that leading whitespaces from tokenization pose a confound to surprisal calculations and that the probability of trailing whitespaces should be included instead. However, we do not include a unit's trailing whitespace in our surprisal calculation; see Giulianelli et al. (2024). Then, following previous work (Meister et al., 2021; Clark et al., 2023), we measure the uniformity of information of a generated completion $\boldsymbol{u}$ given the prefix $\boldsymbol{c}$ using the mean surprisal variance

$$\mathrm{UID}_v(\boldsymbol{u} \mid \boldsymbol{c}) = \frac{1}{N}\sum_{n=1}^{N}(\iota_{\boldsymbol{\theta}}(u_n \mid \boldsymbol{c}_n) - \mu_{\iota_{\boldsymbol{\theta}}}(\boldsymbol{u} \mid \boldsymbol{c}))^2, \tag{22}$$

where the mean surprisal $\mu_{\iota_{\boldsymbol{\theta}}}(\boldsymbol{u} \mid \boldsymbol{c})$ is given by

$$\mu_{\iota_{\boldsymbol{\theta}}}(\boldsymbol{u} \mid \boldsymbol{c}) = \frac{1}{N}\sum_{n=1}^{N}\iota_{\boldsymbol{\theta}}(u_n \mid \boldsymbol{c}_n). \tag{23}$$

Additionally, we calculate the mean local surprisal variance:

$$\mathrm{UID}_{lv}(\boldsymbol{u} \mid \boldsymbol{c}) = \frac{1}{N-1}\sum_{n=2}^{N}\big(\iota_{\boldsymbol{\theta}}(u_n \mid \boldsymbol{c}_n) - \iota_{\boldsymbol{\theta}}(u_{n-1} \mid \boldsymbol{c}_{n-1})\big)^2. \tag{24}$$

To evaluate the diversity of the generations, we calculate the mean unique n-gram ratio (n-Gram%) over completions $\boldsymbol{u}$. In Tab. 11, we report the mean surprisal, surprisal variance, local surprisal variance, and unique n-gram ratios across all data splits. For models fine-tuned without regularization ($\lambda = 0$), surprisal variance and local variance tend to increase compared to the pretrained models, with a few exceptions, particularly for GPT2-M, where the variance and local variance remain close to the pretrained models. Overall, this indicates that information becomes less uniformly distributed. However, under KL regularization ($\lambda = 500$), this trend is reversed, and we observe more uniform information in the generated text with the exception being GPT2-L: Dundee → Dundee. Additionally, we observe a decline in the ratio of unique unigrams compared to the pretrained models, indicating that fine-tuned models generate more repetitive text. However, diversity and uniformity of information are not necessarily linked, as models fine-tuned without regularization tend to generate less diverse and less uniform text.

[14] https://github.com/tpimentelms/probability-of-a-word

Model	↓ $\mu_{\iota_{\boldsymbol{\theta}}}$	↓ UID$_v$	↓ UID$_{lv}$	↑1-Gram%	↑2-Gram%	↑3-Gram%
GPT2-L	3.00	6.69	14.25	84.84	94.16	95.86
GPT2-M	2.94	7.67	16.62	84.39	93.07	94.83
GPT2-S	2.62	5.70	11.65	82.53	91.63	93.35
$\lambda = 0$						
GPT2-L D → D	$3.38_{0.35}$	$10.17_{2.93}$	$20.58_{6.94}$	$67.07_{14.68}$	$74.79_{16.37}$	$76.71_{16.32}$
GPT2-L P → D	$2.96_{0.03}$	$11.29_{1.74}$	$23.76_{4.66}$	$68.88_{1.50}$	$79.78_{1.48}$	$82.79_{1.53}$
GPT2-L Z → D	$3.49_{0.21}$	$9.71_{1.35}$	$20.15_{3.03}$	$78.02_{4.05}$	$92.40_{2.04}$	$95.48_{1.26}$
GPT2-L D → P	$3.03_{0.45}$	$9.40_{2.49}$	$19.26_{4.80}$	$76.41_{3.92}$	$85.80_{4.50}$	$88.17_{4.39}$
GPT2-L P → P	$2.97_{0.04}$	$11.27_{1.74}$	$23.81_{4.64}$	$68.91_{1.76}$	$79.91_{1.66}$	$82.88_{1.65}$
GPT2-L Z → P	$3.71_{0.27}$	$11.22_{2.87}$	$23.38_{5.82}$	$86.74_{0.79}$	$96.36_{0.42}$	$97.98_{0.49}$
GPT2-L D → Z	$3.24_{0.20}$	$7.72_{1.57}$	$14.48_{4.86}$	$61.65_{19.46}$	$68.38_{21.10}$	$70.16_{21.20}$
GPT2-L P → Z	$2.97_{0.03}$	$6.60_{0.08}$	$14.06_{0.19}$	$82.93_{1.91}$	$92.89_{1.27}$	$95.00_{0.86}$
GPT2-L Z → Z	$3.99_{0.12}$	$13.85_{0.98}$	$27.72_{2.46}$	$84.60_{1.91}$	$94.71_{2.05}$	$96.39_{1.68}$
GPT2-M D → D	$3.27_{0.03}$	$8.94_{0.25}$	$18.22_{0.71}$	$80.82_{0.35}$	$91.37_{0.63}$	$93.82_{0.73}$
GPT2-M P → D	$3.04_{0.07}$	$7.81_{0.47}$	$15.85_{1.21}$	$75.82_{0.24}$	$87.77_{0.17}$	$90.91_{0.26}$
GPT2-M Z → D	$3.27_{0.24}$	$7.95_{0.92}$	$16.70_{1.74}$	$83.37_{1.59}$	$93.87_{1.44}$	$95.81_{1.15}$
GPT2-M D → P	$2.83_{0.09}$	$6.96_{1.02}$	$14.73_{2.40}$	$81.05_{1.78}$	$90.15_{1.62}$	$92.25_{1.45}$
GPT2-M P → P	$2.99_{0.04}$	$7.81_{0.46}$	$14.52_{1.42}$	$67.51_{0.78}$	$78.71_{0.55}$	$82.19_{0.68}$
GPT2-M Z → P	$3.08_{0.09}$	$7.25_{0.60}$	$15.18_{1.32}$	$84.68_{0.70}$	$93.70_{0.89}$	$95.37_{0.81}$
GPT2-M D → Z	$2.89_{0.02}$	$7.61_{0.80}$	$15.62_{1.20}$	$81.18_{0.69}$	$91.21_{0.98}$	$93.46_{1.00}$
GPT2-M P → Z	$2.92_{0.03}$	$7.45_{0.45}$	$15.53_{0.79}$	$76.63_{4.35}$	$87.13_{3.91}$	$89.82_{3.52}$
GPT2-M Z → Z	$3.78_{0.13}$	$12.12_{1.42}$	$24.12_{2.59}$	$84.95_{0.86}$	$94.30_{0.85}$	$96.00_{0.80}$
GPT2-S D → D	$2.72_{0.13}$	$6.27_{0.52}$	$13.09_{1.52}$	$81.14_{0.95}$	$89.13_{1.19}$	$91.02_{1.22}$
GPT2-S P → D	$2.94_{0.05}$	$7.03_{0.31}$	$13.92_{0.58}$	$71.91_{2.60}$	$83.59_{2.19}$	$86.61_{1.83}$
GPT2-S Z → D	$2.89_{0.13}$	$7.24_{0.79}$	$14.02_{1.19}$	$79.21_{2.10}$	$89.31_{1.04}$	$91.61_{0.73}$
GPT2-S D → P	$2.72_{0.10}$	$6.54_{0.27}$	$13.05_{0.59}$	$77.08_{3.04}$	$85.98_{2.90}$	$88.30_{2.66}$
GPT2-S P → P	$2.95_{0.07}$	$6.91_{0.25}$	$13.79_{0.93}$	$72.17_{1.99}$	$83.91_{1.42}$	$86.87_{1.19}$
GPT2-S Z → P	$2.62_{0.04}$	$5.49_{0.32}$	$11.21_{0.49}$	$81.92_{1.03}$	$90.86_{0.94}$	$92.63_{0.76}$
GPT2-S D → Z	$2.54_{0.07}$	$5.81_{0.25}$	$11.25_{0.30}$	$76.73_{2.73}$	$85.56_{2.60}$	$87.84_{2.33}$
GPT2-S P → Z	$2.58_{0.01}$	$5.23_{0.27}$	$10.59_{0.73}$	$77.17_{2.97}$	$87.90_{2.13}$	$90.46_{1.60}$
GPT2-S Z → Z	$2.96_{0.13}$	$7.79_{1.02}$	$15.40_{2.02}$	$79.61_{1.94}$	$89.08_{2.34}$	$91.21_{2.32}$
$\lambda = 500$						
GPT2-L D → D	$3.01_{0.04}$	$8.05_{0.77}$	$16.54_{1.34}$	$82.38_{0.18}$	$92.39_{0.40}$	$94.26_{0.43}$
GPT2-L P → D	$2.48_{0.03}$	$5.23_{0.40}$	$10.35_{0.65}$	$76.24_{0.66}$	$86.37_{0.62}$	$88.85_{0.60}$
GPT2-L Z → D	$2.70_{0.02}$	$5.64_{0.01}$	$12.00_{0.10}$	$80.91_{0.63}$	$91.11_{0.39}$	$93.15_{0.33}$
GPT2-L D → P	$2.85_{0.01}$	$5.97_{0.07}$	$12.62_{0.16}$	$83.48_{0.60}$	$92.49_{0.63}$	$94.23_{0.63}$
GPT2-L P → P	$2.46_{0.03}$	$5.14_{0.35}$	$9.95_{0.55}$	$76.75_{0.79}$	$86.68_{0.50}$	$89.01_{0.38}$
GPT2-L Z → P	$2.82_{0.05}$	$6.12_{0.31}$	$13.06_{0.72}$	$83.11_{0.83}$	$92.96_{0.61}$	$94.75_{0.52}$
GPT2-L D → Z	$2.88_{0.03}$	$6.37_{0.19}$	$13.64_{0.41}$	$83.04_{0.71}$	$91.95_{0.71}$	$93.69_{0.63}$
GPT2-L P → Z	$2.34_{0.01}$	$4.46_{0.07}$	$8.84_{0.09}$	$73.86_{0.99}$	$84.26_{0.90}$	$86.96_{0.84}$
GPT2-L Z → Z	$2.75_{0.02}$	$5.61_{0.09}$	$11.96_{0.30}$	$81.12_{0.20}$	$91.51_{0.42}$	$93.63_{0.48}$
GPT2-M D → D	$2.85_{0.02}$	$6.63_{0.35}$	$13.98_{0.78}$	$83.22_{1.08}$	$92.17_{0.33}$	$94.09_{0.14}$
GPT2-M P → D	$2.70_{0.03}$	$5.73_{0.54}$	$11.74_{1.14}$	$79.41_{0.88}$	$90.14_{0.48}$	$92.64_{0.31}$
GPT2-M Z → D	$2.72_{0.04}$	$6.07_{0.17}$	$12.76_{0.30}$	$81.52_{0.78}$	$91.55_{0.64}$	$93.72_{0.56}$
GPT2-M D → P	$2.80_{0.07}$	$6.79_{0.83}$	$14.41_{2.03}$	$83.00_{1.39}$	$91.46_{1.04}$	$93.41_{0.91}$
GPT2-M P → P	$2.46_{0.03}$	$4.54_{0.17}$	$9.16_{0.42}$	$75.91_{0.85}$	$86.71_{0.59}$	$89.65_{0.56}$
GPT2-M Z → P	$2.72_{0.03}$	$6.00_{0.04}$	$12.59_{0.10}$	$81.54_{1.14}$	$91.27_{0.78}$	$93.28_{0.76}$
GPT2-M D → Z	$2.80_{0.06}$	$6.75_{0.14}$	$14.05_{0.34}$	$81.90_{1.04}$	$91.21_{0.82}$	$93.18_{0.80}$
GPT2-M P → Z	$2.57_{0.09}$	$5.32_{0.46}$	$10.82_{0.98}$	$77.08_{1.73}$	$88.04_{1.19}$	$90.78_{0.98}$
GPT2-M Z → Z	$2.72_{0.03}$	$5.88_{0.05}$	$12.17_{0.21}$	$81.94_{0.91}$	$91.53_{0.55}$	$93.52_{0.51}$
GPT2-S D → D	$2.37_{0.09}$	$4.50_{0.11}$	$9.30_{0.17}$	$76.56_{2.02}$	$85.95_{2.28}$	$87.90_{2.23}$
GPT2-S P → D	$2.13_{0.08}$	$4.30_{0.28}$	$8.62_{0.70}$	$68.97_{2.33}$	$78.38_{2.49}$	$81.41_{2.41}$
GPT2-S Z → D	$2.34_{0.02}$	$4.37_{0.05}$	$8.84_{0.13}$	$74.31_{1.18}$	$85.19_{0.79}$	$87.66_{0.66}$
GPT2-S D → P	$2.40_{0.08}$	$4.60_{0.23}$	$9.69_{0.43}$	$79.17_{2.12}$	$87.65_{1.83}$	$89.28_{1.69}$
GPT2-S P → P	$2.15_{0.04}$	$4.11_{0.10}$	$8.34_{0.19}$	$69.54_{0.43}$	$79.13_{0.51}$	$82.08_{0.50}$
GPT2-S Z → P	$2.45_{0.05}$	$5.07_{0.36}$	$10.47_{0.57}$	$79.42_{0.94}$	$88.60_{0.93}$	$90.52_{0.79}$
GPT2-S D → Z	$2.36_{0.05}$	$4.52_{0.12}$	$9.36_{0.29}$	$77.73_{1.12}$	$86.73_{1.04}$	$88.58_{1.02}$
GPT2-S P → Z	$2.17_{0.11}$	$4.22_{0.20}$	$8.65_{0.46}$	$70.47_{2.41}$	$79.94_{2.76}$	$82.95_{2.69}$
GPT2-S Z → Z	$2.35_{0.03}$	$4.97_{0.18}$	$10.25_{0.47}$	$75.92_{0.68}$	$85.66_{0.77}$	$87.76_{0.88}$

Table 11: Full evaluation results for completions generated on prefixes sampled from CNN/Dailymail.