Eye Movement Features Can Predict Human Preferences on Machine-Generated Texts

Anonymous ACL submission

Abstract

Eye-tracking metrics offer valuable insights into human visual attention during language comprehension, yet existing research and resources in this area are limited. To bridge this gap, we introduce ***G**aze **R**esponses for **E**valuating **A**I **T**exts* (GREAT), a comprehensive dataset capturing human eye-movement patterns during screen reading of passages generated by large language models (LLMs). The dataset includes raw eye-movement recordings, reading-time measures, and post-reading evaluations for LLM-generated passage pairs selected from MT-Bench dataset, alongside rigorous validation metrics. The collected eye-tracking metrics demonstrate strong explanatory power in predicting text quality. When integrated with negative log-likelihood (NLL), a commonly used metric for evaluating text quality, it substantially enhances model performance across all standard statistical criteria. These findings demonstrate that eye-tracking data effectively complement probabilistic metrics, improving predictive accuracy for text quality assessment. The full dataset and some processing code are publicly available at `https://anonymous.4open.science/r/eye-track`.

1 Introduction

Understanding how humans perceive and evaluate machine-generated text is a growing area of research in natural language processing (NLP), especially as large language models (LLMs) become increasingly integrated into real-world applications. Despite progress in automatic metrics—from n-gram-based scores like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) to model-based ones like BERTScore (Zhang et al., 2019) and BLEURT (Sellam et al., 2020)—they often miss human preference nuances, while human evaluations, though reliable, are costly and inconsistent. This highlights the need for scalable, cognitively grounded alternatives.

Eye-tracking has long been established as a robust method in psycholinguistics for studying cognitive processing during reading. Metrics such as fixation duration, saccade frequency, and regression behavior (backward saccades) offer real-time insights into a reader's attention, effort, and comprehension. These metrics have been extensively validated as indicators of text difficulty and are linked to theoretical constructs like surprisal and information density (Smith and Levy, 2013; Meister et al., 2021; De Varda and Marelli, 2023; Shain et al., 2024). However, their application to the evaluation of machine-generated text—especially from modern LLMs—remains relatively underexplored.

Naturally, some recent endeavors have started exploring the relationship between eye-movement and LLM-generated texts (Bolliger et al., 2024). However, still little is known about the strength of the relations (Oh and Schuler, 2022). It is not clear whether the eye-movement metrics can directly reflect the quality of model-generated texts, whether they are correlated with human judges' preferences, and to what extent the metrics can be used as a way for evaluation (Lopez-Cardona et al., 2025).

In this study, we focus on studying the end users' instantaneous eye-movement reaction to the model-generated texts through a comprehensive experimental investigation that bridges psycholinguistic methods with modern NLP evaluation. We introduce GREAT (Gaze Responses for Evaluating AI Texts), a dataset that captures eye-movement data from human readers as they read and evaluate LLM-generated responses. Based on the MT-Bench dataset, which provides human preference labels for pairs of LLM-generated texts, the collected dataset includes not only gaze data but also reading times, fixation patterns, and post-reading quality judgments. By systematically analyzing this data, we aim to uncover how various aspects of reading behavior—both temporal and spatial—can serve as proxies for human assessments of text quality.

Our central research question is: To what extent can gaze-based features predict the perceived quality of LLM-generated text, especially when compared to or combined with model-based metrics such as NLL? To explore this, we evaluate the predictive power of several eye-tracking features—fixation time, pixel dwelling time, and backward saccade frequency—and assess how well these features align with human preferences in the MT-Bench evaluation framework. To sum up, our work offers the following two major contributions:

- **A novel eye-tracking dataset (GREAT)** that captures rich, fine-grained gaze behavior from participants reading LLM-generated text pairs.
- **Validation:** We demonstrate that eye-tracking metrics significantly enhance the predictive power of text quality assessment when combined with traditional measures like negative log-likelihood (NLL).

2 Experiment Setup

2.1 Textual materials

Our study is enabled by the MT-Bench and Chatbot Arena dataset (henceforth MT-bench) (Zheng et al., 2023). MT-Bench is an open-ended question-answer dataset created for evaluating chatbots' conversational skills and instruction-following capabilities, comprising 30,000 machine-generate conversations with human preference annotations to support further research. We meticulously curated a targeted subset of plain content from this dataset to serve as the reading materials, guided by the following principles:

Text length To prevent scrolling or page flipping, it is essential that the texts presented to participants fit entirely within a single screen. The average length of the final selected texts is accordingly set to 65 words.

Text domain To accommodate the diverse backgrounds of the participants, we exclude text materials related to mathematics and programming code, and retain only those written in plain English, with their source from Wikipedia, news articles, etc. This ensures that the majority of subjects can understand the material without difficulty caused by domain knowledge.

With the above selection standards, we aim to investigate how the quality and complexity of textual content influence visual attention and reading behavior. This approach also allowed us to explore whether metrics derived from machine judgment correlate with gaze-based indicators of text quality, thereby linking the subjective experience of reading with objective model evaluations.

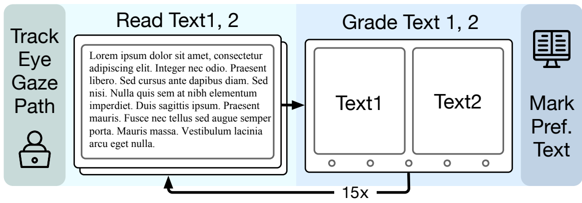


Figure 1: **Task workflow for dataset construction.** Participants completed 15 sessions, each involving the sequential reading of two texts while their eye gaze was recorded. After reading, participants rated their relative preference for the texts using a five-point scale displayed beneath the text pairs on the grading interface.

2.2 Data Collection

2.2.1 Participants

A total of 38 participants (10 female; mean age = 20.76 years, SD = 1.96), all enrolled in bachelor's or master's programs with full-time English training, took part in the eye-tracking experiment. Before the experiment, each participant was required to complete a questionnaire assessing their English proficiency. An overview of the participants' self-reported proficiency levels is provided in Appendix B, which is compared with the typical score bands from various common English proficiency tests (e.g., TOEFL, IELTS, CET-4/CET-6) to contextualize these levels (refer to Table 4).

2.2.2 Apparatus

The experiment was conducted in a relatively controlled environment, and unrelated personnel were cleared during the experiment. Set up screens and provide noise-cancelling headphones to reduce external distractions. Before the start of the experiment, the participants were asked to complete the calibration, ensuring the effectiveness of the collected data. During the experiment, the participants were instructed to remain immobile. The display screen used in the experiment is 28 inches, and the resolution is 2560×1440 pixels. Eye movement data were recorded using a commercial eye tracker, operating at a sampling rate of 60 Hz.

2.2.3 Procedure

We recruited 38 participants for our eye-tracking reading experiment. Each participant watched a pre-recorded demonstration video explaining the experimental procedure. Participants were instructed to read the text at their own pace without being tested on comprehension and to remain as still as possible. To ensure accurate eye-tracking data collection, each participant has completed a six-point calibration procedure before the given task.

Participants used custom keyboards during the experiment to minimize distractions, prevent accidental touches, and reduce data noise caused by looking for the keyboard. The reading process was divided into two stages: reading and rating. In the reading stage, participants pressed the "start" button to display the text and the "end" button to conclude the reading session before proceeding to the next phase. Each participant completed two reading sessions before moving on to a rating task, in which they used a five-point Likert scale (Likert, 1932) to indicate their preference for the two texts they had just read. To maintain a high level of attention, participants were required to complete their preference selection within 30 seconds. Each subject completed 15 cycles of this reading and rating process.

To ensure the authenticity of the preference selection, the texts being compared were displayed on the screen. Participants selected their preferred text from the options at the bottom of the screen, followed by a confirmation pop-up. To prevent accidental submissions, a throttle measure was implemented, ensuring that the submission could not be made more than once within one second.

3 Data Processing

3.1 Data Cleaning

To enhance statistical validity, samples with total reading duration outside the range of the mean plus or minus two standard deviations were excluded. Additionally, samples with insufficient valid gaze points were removed to control for technical artifacts and ensure the data accurately represent participants' cognitive behavior. In addition, data that was not rated in a timely manner was deleted. With such a criterion, 83 pairs of reads were excluded, resulting in 487 pairs of reading data retained.

The eye tracker used in our experiment records binocular data, capturing gaze positions from both the left and right eyes. To reduce systematic error and enhance the accuracy of gaze estimation, we compute the average of the left and right eye position signals, following established practices in eye-tracking methodology (Hooge et al., 2019).

The Area of Interest (AOI) is defined as the bounding box of a word (distinguished by a space), with punctuation marks incorporated into the preceding word (Holmqvist and Andersson, 2017; Hessels et al., 2016; Hooge et al., 2025). By setting the boundary of the reading content, the gaze points located outside the expected reading area (e.g., the edges of the screen) are excluded. Remaining points will be assigned to the nearest AOI.

A common issue in eye-tracking experiments is vertical drift, characterized by a gradual displacement of the recorded gaze coordinates over time (Carr et al., 2022; Chen et al., 2021; Frank and Aumeistere, 2024). It can be resolved by clustering-based classification approaches (e.g., `AgglomerativeClustering` in scikit-learn).

To detect microsaccades, we employed a velocity-based algorithm (Engbert and Kliegl, 2003; Nyström and Holmqvist, 2010), which identifies saccadic events as velocity outliers relative to the overall distribution. To minimize noise and enhance signal stability, eye movement data were first smoothed using a five-sample weighted moving average. Assuming an approximately normal distribution of velocity values, the detection threshold was defined as two standard deviations above the mean velocity across all samples. This approach has been shown to offer robust and consistent performance in identifying microsaccadic activity.

To analyze the reading trajectory, we adopt a time-space-based clustering method known as Spatial Temporal-DBSCAN (Birant and Kut, 2007). By setting temporal and spatial thresholds, the fixation points are clustered in time and space, and certain noisy data is excluded. The clusters are assigned to the corresponding AOI regions to analyze the scan paths. The scanning path over the area of interest (AOI) is illustrated in Figure 4; for additional details, refer to Appendix C.

3.2 Dataset Overview

In this study, the dataset we introduced captures detailed behavioral traces of readers during the reading process by these attributes:

Reading Time: Time for reading a text. The average reading time in the dataset is 21369.07 ms (SD $= 9345.94$). A maximum of 12.51% of reading time is concentrated within the interval [17756, 20756].

Pixel Dwelling Time (PDT): Average eye movement time per pixel. The average PDT in the dataset is 1.9 ms per pixel (SD $= 1.11$). A maximum of

21.48% of PDT is concentrated within the interval [1.34, 1.84]. The data volume around 4.5 ms per pixel gradually approaches zero.

Saccade Frequency (SF): Ballistic movements, the eye rapidly shifts its focus from one fixation point to another. The average number of saccades is 109.10 (SD $= 45.70$). A maximum of 17.16% of SF is concentrated within the interval [97, 117].

Backward Saccade Frequency (BSF): Backward Saccade refers to the reader moving their eyes backward to the text they have previously read. The average number of backward saccades is 48.44 (SD = 18.64). A maximum of 32.48% of PDT is concentrated within the interval [46, 61].

Fixation Time (FT): The definition of fixation time is the total reading duration minus the saccade time. The average fixation time is 18614.25 ms (SD $=$ 8572.29). A maximum of 13.57% of PDT is concentrated within the interval [19528, 22528].

Negative Log-Likelihood (NLL) quantifies the uncertainty of a language model by measuring the negative log-probability it assigns to each word given its preceding context. Formally, for a sequence of word–context pairs (x_i, y_i), NLL is defined as:

$$\mathcal{L}(\theta) = -\sum_{i=1}^{N} \log p\left(y_i \mid x_i; \theta\right)$$

where $p(y_i \mid x_i; \theta)$ is the model's predicted probability. In this study, NLL was computed using the LLaMA-7B model (Touvron et al., 2023) to provide a consistent estimate of token-level uncertainty across different texts. Notably, GPT-4-generated texts displayed the narrowest NLL distribution, reflecting stable predictive confidence, while Claude-v1 (Anthropic, 2023) outputs exhibited a broader, more skewed distribution with occasional high NLL values, suggesting greater variability in prediction difficulty.

MT-bench Score (MT) & Experiments score (EXP): The score of text materials from MT-bench (1.0, 1.5, 2.0) and the score of text materials during experiments (1.0, 1.25, 1.5, 1.75, 2.0). The majority scores of MT demonstrate clear preferences between text pairs. A similar pattern is observed in the EXP results. The texts from the gpt-3.5-turbo (OpenAI, 2023) model are the largest, and its distribution covers all MT and EXP values, especially the proportion of MT = 2.0 and EXP = 2.0 is relatively high. The GPT-4 model (OpenAI et al., 2024) has the least amount of text data and a relatively uniform distribution of ratings. The LLaMA-13B (Touvron et al., 2023) model has a high proportion of text generation on MT = 1.0 and EXP = 1.0.

Appendix D illustrates more detailed statistics about the dataset.

3.3 Dataset Validation

To assess the reliability and validity of our dataset, we conducted both benchmark-based and correlation-based validation analyses. Specifically, we evaluated the agreement between participants' experimental preference scores and the standardized MT-Bench scores. After aligning the scoring scales for comparability—where scores of 1.25 and 1.75 were mapped to 1.0 and 2.0, respectively, while all other values remained unchanged—the matching accuracy reached 80%, consistent with prior findings reported by Zheng et al. (Zheng et al., 2023). This level of agreement provides strong support for the methodological robustness and external validity of our experimental framework.

Additionally, we investigated the relationship between fixation duration and the surprisal of texts, which is defined as:

$$\text{Surprisal}(w_t) = -\log p\left(w_t \mid w_1, \ldots, w_{t-1}\right)$$

captures the notion that less predictable words are cognitively more demanding and are therefore associated with increased reading times (Wilcox et al., 2025). To empirically evaluate this relationship, we computed Pearson and Spearman correlation coefficients between surprisal values and fixation durations across the dataset (Mukaka, 2012). The analysis revealed weak but statistically significant positive correlations ($\rho_{\text{Pearson}} = 0.107$, $\rho_{\text{Spearman}} = 0.072$, both at $p < 0.001$ level). This is consistent with predictions derived from established cognitive models of reading, such as E-Z Reader (Reichle et al., 1998) and SWIFT (Engbert et al., 2002).

While the positive association supports the theoretical link between lexical predictability and reading behavior, the relatively low magnitude of the correlation indicates that a broader set of factors likely influences fixation duration. These may include individual cognitive differences, reading strategies, task demands, and syntactic or discourse-level complexity (Sheridan and Reichle, 2016). Together, these results suggest that the dataset captures psycholinguistically sound variance while also reflecting the multifactorial nature of human reading behavior.

4 Experiments

Building upon the demonstrated utility of our dataset for analyzing reading behavior, this section investigates the nuanced influence of eye gaze features on human preferences when evaluating LLM-generated texts. Grounded in psycholinguistic research, eye-tracking metrics offer a valuable window into the non-cognitive dimensions of reading, reflecting the intricate interplay of attention, cognitive load, and comprehension processes within human-machine interaction contexts. To rigorously examine these relationships, we employ a series of linear mixed-effects models. This statistical framework was chosen to enable a detailed examination of the effects of linguistic features on eye-tracking variables.

4.1 Multicollinearity analysis and variable selection

The limited proportion of variance in the outcome measure explained by language model identity suggests that the disparities observed between different models are primarily inherent to their architecture and training, rather than significantly contributing to the explanation of variation in our target variable. Consequently, to more effectively isolate the effects of eye-tracking and linguistic features as potential indicators of underlying cognitive processing, language model identity was excluded from subsequent analyses.

To address multicollinearity among candidate predictors for MT-score prediction, we use Variance Inflation Factors (VIF) (Toothaker, 1994) as a diagnostic tool to quantify how much the variance of each predictor is inflated due to correlations with other predictors, guiding feature selection toward independent and interpretable variables. We retain predictors with VIF values below 5, indicating acceptable levels of multicollinearity.

Predictor	**VIF ↓**
Experiment Score	1.010692
Negative Log-Likelihood (NLL)	1.195635
Fixation Time (FT)	1.335048
Pixel Dwelling Time (PDT)	1.832083
Backward Saccade Frequency (BSF)	2.255633

Table 1: The predictors with small VIF values.

Predictors exhibiting VIF values exceeding this threshold would typically be considered for exclusion or further investigation to mitigate the adverse effects of high intercorrelation. Based on Table 1, we include NLL (linguistic predictor) and three eye-tracking metrics—PDT, BSF, and FT—as predictors in subsequent models.

4.2 Hypotheses

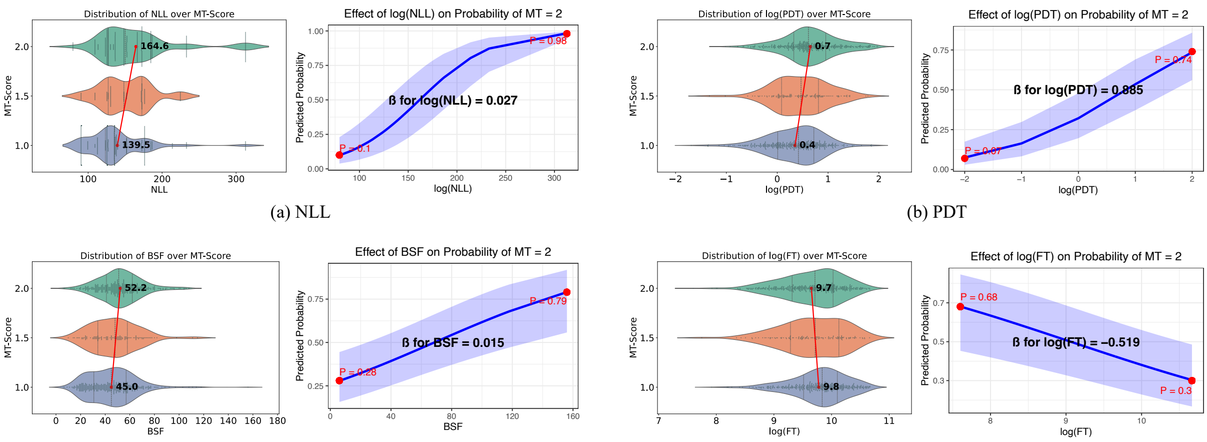
We propose the following hypotheses regarding the relationship between eye-tracking variables and perceived text quality (measured via MT-score):

- **H1**: Lower Pixel Dwelling Time (faster reading speed) positively predicts perceived text quality, reflecting smoother and more fluent reading behavior.
- **H2**: Increased Backward Saccade Rate positively predicts text quality, indicating more frequent regressions during reading are associated with better comprehension, deeper processing, or higher-quality writing.
- **H3**: Longer Fixation Time is negatively associated with quality judgments as it implies that the comprehension process demands greater cognitive effort.
- **H4**: Combining temporal (Fixation Time), spatial (Pixel Dwelling Time), and difficulty (Backward Saccade Frequency) eye-tracking metrics improves predictive accuracy beyond either metric individually, demonstrating their complementary contributions.

We posit that gaze-based measures offer a richer and more direct window into readers' cognitive engagement with text than purely statistical linguistic features. Specifically, these metrics provide a multidimensional characterization of reading behavior: Fixation Time captures processing effort, Pixel Dwelling Time reflects reading efficiency, and Backward Saccade Frequency indicates comprehension difficulty. These features serve as key predictors in our analysis, enabling a detailed examination of how different aspects of reading behavior manifest in response to text quality.

4.3 Mixed-effect linear models

This section presents mixed-effects regression models examining the relationship between human-judged text quality scores (MT-Score) and predictors–NLL and eye-tracking metrics including PDT, BSF, and FT. Random intercepts account for variability across text generators and participants. The findings offer strong support for the earlier hypotheses, showing that both cognitive processing



(a) NLL (b) PDT

(c) BSF (d) FT

Figure 2: Distributions of the four main predictors, NLL, PDT, BSF, and FT, grouped by MT-Score values (1.0, 1.5, and 2.0). Red dots on the left represent the median of each distribution. The four corresponding probability curves (right blue) show the effect of four main predictors on the probability of MT $= 2.0$, along with the regression coefficients (shadow areas are 95% confidence intervals).

and model uncertainty are significant predictors of perceived text quality.

Model	β	SE	t-value	Effect Direction
M_{NLL}	2.699×10^{-2}	3.174×10^{-2}	8.504	Positive
M_{PDT}	8.850×10^{-1}	1.467×10^{-1}	6.034	Positive
M_{BSF}	1.520×10^{-2}	4.163×10^{-3}	3.651	Positive
M_{FT}	-5.188×10^{-1}	1.477×10^{-1}	-3.514	Negative

Table 2: Parameter estimates, standard errors, t-values, and model selection criteria (AIC, BIC) for mixed-effects models predicting MT-Score from individual predictors. Random intercepts for both *Generator* and *Subject* were incorporated in all single predictor models, except for M_{NLL}, which included a random intercept for *Generator* only.

Base Model (M_{NLL}) As a foundational model, MNLL examines the effect of token-level uncertainty, measured via NLL, on perceived text quality. The model is specified as:

$$\text{MT} \sim \beta_0 + \beta_{\text{NLL}} \cdot \text{NLL} + (1|\text{Generator}) + \epsilon$$

, where β_0 is the intercept, $(1|\text{Generator})$ is a random effect accounting for variability across different text generators from the MT-Bench dataset, and ϵ is residual error.

The coefficient for NLL is positive and highly significant (see Table 2), indicating that texts with higher token-level surprisal tend to receive higher MT-Scores. This finding corroborates prior work suggesting that readers favor content that is more novel or informative (Gehrmann et al., 2019), and it establishes a benchmark for evaluating the added predictive value of eye-tracking measures.

Model 1: Effect of Pixel Dwelling Time (M_{PDT}) PDT as a measure of reading time can reflect the average duration of visual attention for a reading session (Rayner, 1998). We fit a mixed-effect linear model with name M_{PDT}, formulated as:

$$\begin{aligned}\text{MT} \sim\ & \beta_0 + \beta_{\log(\text{PDT})} \cdot \log(\text{PDT}) \\ & + (1|\text{Generator}) + (1|\text{Subject}) + \epsilon\end{aligned}$$

The formula includes random intercepts for both the generator and the subject. Results show a statistically significant positive effect of PDT on MT-Score (see Table 2), indicating that longer visual attention is associated with higher subjective quality ratings. This finding supports hypothesis **H1**, indicating that increased cognitive engagement, as reflected by PDT, corresponds to more favorable human evaluations.

Model 2: Effect of Backward Saccade Frequency (M_{BSF}) BSF captures the rate of regressions—eye movements returning to earlier parts of the text—typically linked to increased cognitive effort during reading, such as resolving ambiguity or reprocessing complex content. To evaluate its relationship with perceived text quality, we estimated a mixed-effects linear model (M_{BSF}) defined as:

$$\begin{aligned}\text{MT} \sim\ & \beta_0 + \beta_{\text{BSF}} \cdot \text{BSF} \\ & + (1|\text{Generator}) + (1|\text{Subject}) + \epsilon\end{aligned}$$

Results show a statistically significant positive effect of BSF on MT-Score (see Table 2), indicating that texts eliciting more backward saccades tend to receive higher human quality ratings. This supports hypothesis **H2**, suggesting that readers engage more deeply—and perhaps more favorably—with texts that prompt increased rereading behavior.

Model 3: Effect of Fixation Time ($\mathbf{M_{FT}}$) FT measures the cumulative duration of a participant's gaze fixations on a given text, serving as an indicator of processing effort during reading. To investigate its role in predicting perceived text quality, we specified M_{FT} as follows:

$$\begin{aligned} \mathrm{MT} \sim & \beta_0 + \beta_{\log(\mathrm{FT})} \cdot \log(\mathrm{FT}) \\ & + (1|\mathrm{Generator}) + (1|\mathrm{Subject}) + \epsilon \end{aligned}$$

The model reveals a significant negative relationship between FT and MT-Score (see Table 2), indicating that shorter fixation durations are associated with higher subjective quality ratings. This finding supports hypothesis **H3**, suggesting that more easily processed texts, reflected in reduced fixation time, are perceived as higher in quality.

The distribution and regression results of the base model and all three single predictor models are shown in Figure 2.

Multi-Predictor Models Integrating multiple eye-tracking predictors can capture complementary aspects of reading behavior that jointly improve the assessment of text quality (Mathias et al., 2018).

Model	AIC	BIC	R^2	Adj. R^2
		Base Model		
M_{NLL}	1003	1017	6.182×10^{-2}	6.085×10^{-2}
		Single Predictor Models		
M_{PDT}	1067	1086	6.436×10^{-2}	6.340×10^{-2}
M_{FT}	1094	1113	1.343×10^{-2}	1.242×10^{-2}
M_{BSF}	1093	1112	3.255×10^{-2}	3.155×10^{-2}
		Multi-Predictor Models		
M_{PDT+FT}	1042	1066	9.122×10^{-2}	8.935×10^{-2}
$M_{PDT+BSF}$	1069	1093	6.666×10^{-2}	6.474×10^{-2}
M_{FT+BSF}	1069	1093	6.381×10^{-2}	6.188×10^{-2}
M_{EYE}	1040	1069	9.924×10^{-2}	9.645×10^{-2}
$M_{EYE+NLL}$	**982.7**	**1016**	$\mathbf{11.91 \times 10^{-2}}$	$\mathbf{11.55 \times 10^{-2}}$

Table 3: Model Performance Comparison: AIC, BIC, R^2, and Adjusted R^2 for models predicting text quality grading (MT) using individual and combined metrics. EYE denotes the combination of three eye-tracking metrics—PDT, FT, and BSF—for brevity.

To evaluate the predictive capacity of various models for text quality grading, in Table 3 we compared their performance across four key metrics: Akaike Information Criterion (AIC) (Bozdogan, 1987), Bayesian Information Criterion (BIC) (Schwarz, 1978), coefficient of determination (R^2), and adjusted R^2 (R^2_{adj}). AIC and BIC measure model quality by balancing goodness-of-fit with complexity, where lower values indicate more parsimonious models (Lehtonen et al., 2019). R^2 quantifies the proportion of variance explained by the predictors, while R^2_{adj} adjusts for model complexity, allowing for fairer comparisons between models with differing numbers of features.

Relative to the base model M_{NLL}, which uses only NLL as a predictor, all two-predictor combinations involving eye-tracking metrics (e.g., M_{PDT+FT}, M_{FT+BSF}) demonstrate improved explanatory power across all evaluation metrics. This confirms that gaze-derived behavioral features provide meaningful information beyond token-level model uncertainty (Wiechmann et al., 2022). Notably, the M_{PDT+FT} model achieves the lowest AIC and BIC along with the highest R^2 and R^2_{adj}, indicating the most favorable trade-off between model fit and complexity among the combinations.

M_{EYE}, which integrates the effects of PDT, FT, and BSF, shows a modest improvement over M_{PDT+FT}, though it yields a slightly higher BIC, indicating a minor trade-off in model parsimony. By combining the eye-tracking metrics and NLL, When the eye-tracking metrics are combined with NLL, the model $M_{EYE+NLL}$ achieves the best performance across all four evaluation criteria—AIC, BIC, R^2 R^2_{adj}. The substantial increase in both goodness-of-fit measures suggests that eye-tracking features serve as a valuable complement to NLL in predicting text quality grading, offering robust empirical support for **H4**.

4.4 Correlations with other variables

Model name, a categorical factor representing different language models (e.g., Claude-v1, GPT-3.5-Turbo, GPT-4, LLaMA-13B, Vicuna-13B-v1.2 (Chiang et al., 2023), and Alpaca-13B), exhibits high correlation with mt-scores but is excluded from predictive models to avoid confounding. This is due to its nature as a label rather than a mechanistic predictor, impairing interpretability regarding why different models yield distinct MT scores (Gelman and Hill, 2006; Shmueli, 2010). Moreover, linear models including `model_name` yielded higher R-squared values, indicating that model identity strongly predicts MT-score. To isolate the effects of eye gaze variables within a consistent model context, we exclude `model_name` from the main predictive models, thereby controlling for model generation capabilities and focusing on cognitive processing indicators.

5 Related Works

5.1 Eye Tracking Metrics

In recent years, eye-tracking technology has been widely applied in the study of text readability and comprehension. Here are two important eye-tracking indicators, **fixations** and **backward saccades**. Eye fixations, occurring when the eyes remain relatively still, allow the reader to extract information from the text. Backward saccades, the reader's actions of moving their eyes (regress) back to previously read material in the text (i.e., regression). Generally, research indicates that as text difficulty increases, fixation durations lengthen, the scanning distance shortens, and there are more backward saccades during reading (Rayner, 1998, 2009). In multiple languages, these eye movement features are significantly correlated with traditional readability scores, indicating that eye movement data can effectively reflect the difficulty of reading text (Baazeem et al., 2021, 2025; Atvars and Aigars, 2017). From this, we can infer that the reading **duration** will also increase as the difficulty of the text increases, which will also be a variable studied in this project.

5.2 Text quality assessment

Text quality assessment is a core issue in natural language processing. With the development of deep learning technology, researchers have proposed various evaluation methods, ranging from traditional rule-based metrics, such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), to modern pre-trained model-based evaluation methods, such as BERTScore (Zhang et al., 2019) and BLEURT (Sellam et al., 2020). Among these language model-based evaluation methods, NLL is commonly used to describe and quantify the predictability of language and its impact on cognitive effort, especially as it can more accurately reflect the difficulty of reading compared to traditional readability formulas (Klein et al., 2025). (Smith and Levy, 2013) showed that NLL is directly proportional to reading time. The higher the NLL of a word, the longer the reader will stay on the word.

5.3 Human judgements of generated texts

Using human judges as a source of evaluation for model-generated texts is a well-practiced (Celikyilmaz et al., 2020), which is one of the three major text generation evaluation methods. However, there is a great variation in the way humans assess it. Hashimoto et al. (2019) points out the shortcomings of human evaluation scores – the lack of preferences over diverse texts. The article (van der Lee et al., 2019) highlights the importance of using multiple raters in the evaluation process. The article (van der Lee et al., 2021) emphasizes the selection of a sample that reflects the target audience and considers how to reduce the sequential effect.

MT-Bench dataset (Zheng et al., 2023) is used as the experimental material in this study. The original dataset is a collection of multiple language models' responses to a set of prompts that come in pairs, covering multiple domains, along with human judges' preferences for each pair (which one wins over the other).

Based on existing research, our study chooses to use LLM-generated text, combining the two eye movement variables, duration and back, and NLL, with human evaluations, to investigate the quality of the model's generated text.

6 Conclusions

In this study, we introduced the GREAT dataset, a useful resource for analyzing human cognitive responses to different LLM-generated text content. The dataset comprises a comprehensive set of eye gaze measurements collected from controlled screen reading experiments. Our preliminary analysis of the data demonstrates that eye movement features, such as reading duration, reading speed, forward/backward saccades, and fixation time, are significantly correlated with human judgments of text quality. Among these features, the effect of *reading speed* is the most significant – almost the same level as that of the negative likelihood of texts (or surprisal). Our analysis validates the GREAT dataset as a robust resource for evaluating LLMs using a human-as-judge approach. The GREAT dataset offers new insights into how individuals interact with and evaluate machine-generated content, highlighting the potential of eye-gaze features as objective indicators of text quality.

7 Limitations

While all participants in our study were experienced English users with at least a decade of language exposure, they were not native speakers, which may introduce subtle differences in reading behavior compared to L1 readers. Additionally, the dataset used in our experiment is limited in domain scope, which may affect the generalizability of our findings across different text types or genres. Our focus on English texts further narrows the applicability of the results to other languages, particularly those with distinct linguistic or orthographic characteristics. The experimental setting, while controlled, may also influence natural reading behavior; participants read in a lab environment with tasks that might not fully replicate everyday reading conditions. Future work could address these limitations by including a more diverse participant pool, expanding text types and domains, incorporating multilingual materials, and considering ecologically valid reading settings to support more comprehensive insights into eye gaze behavior.

8 Ethic Statement

All participants in this study were informed of the research objectives, procedures, and their rights prior to enrollment. The following principles guided the ethical conduct of the research:

- **Informed Consent** Participants provided written informed consent after receiving a detailed explanation of the study, including the use of eye-tracking technology, the nature of tasks (reading and rating texts), and the voluntary nature of their participation. They were advised that they could withdraw from the study at any time without penalty.

- **Confidentiality of Personal Information** Personal data, including names, genders, ages, and English proficiency details, were anonymized and stored securely. Access to raw data was restricted to authorized researchers only. Unless explicitly permitted by the participant, no personally identifiable information (PII) was shared with third parties. Government authorities or ethics review committees could access de-identified data for regulatory purposes, in accordance with institutional policies.

- **Data Usage and Security** Eye-tracking recordings, reading time measures, and preference ratings were used solely for research purposes outlined in this study. All data were encrypted during storage and transmission. Participant identities were separated from research data, and only aggregated, anonymized results were reported in publications or presentations.

- **Participant Obligations** Participants were instructed to maintain the confidentiality of experimental materials and procedures. They were explicitly advised not to disclose details about the study (e.g., text content, rating criteria) to third parties to prevent contamination of results.

- **Ethical Review** This study was conducted in compliance with the Declaration of Helsinki and approved by the institutional ethics committee [insert specific committee name if applicable]. All procedures were designed to minimize potential risks and maximize the scientific value of the research. The collected data do not contain any personal information such as name or personal ID.

By adhering to these principles, the research team ensures the protection of participant rights and maintains the integrity of the study.

References

Anthropic. 2023. Claude ai. https://www.anthropic.com/. Accessed: 2025-05-19.

Atvars and Aigars. 2017. Eye movement analyses for obtaining readability formula for latvian texts for primary school. *Procedia Computer Science*, 104:477–484.

Ibtehal Baazeem, Hend Al-Khalifa, and Abdulmalik Al-Salman. 2021. Cognitively driven arabic text readability assessment using eye-tracking. *Applied Sciences*, 11(18).

Ibtehal Baazeem, Hend Al-Khalifa, and Abdulmalik Al-Salman. 2025. Araeyebility: Eye-tracking data for arabic text readability. *Computation*, 13(5).

Derya Birant and Alp Kut. 2007. St-dbscan: An algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering*, 60(1):208–221.

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. *arXiv preprint arXiv:2408.04289*.

Hamparsum Bozdogan. 1987. Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370.

Jon W Carr, Valentina N Pescuma, Michele Furlan, Maria Ktori, and Davide Crepaldi. 2022. Algorithms for the automated correction of vertical drift in eye-tracking data. *Behavior Research Methods*, 54(1):287–310.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Ming Chen, Raymond R Burke, Sam K Hui, and Alex Leykin. 2021. Understanding lateral and vertical biases in consumer attention: An in-store ambulatory eye-tracking study. *Journal of Marketing Research*, 58(6):1120–1141.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Andrea De Varda and Marco Marelli. 2023. Scaling in cognitive modelling: A multilingual approach to human reading times. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 139–149.

Ralf Engbert, André Longtin, and Reinhold Kliegl. 2002. A dynamical model of saccade generation in reading based on spatially distributed lexical processing. *Vision Research*, 42(5):621–636.

Reinhold Engbert and Reinhold Kliegl. 2003. Microsaccades uncover the orientation of covert attention. *Vision Research*, 43(9):1035–1045.

Stefan L Frank and Anna Aumeistere. 2024. An eye-tracking-with-eeg coregistration corpus of narrative sentences. *Language Resources and Evaluation*, 58(2):641–657.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. 2019. Gltr: Statistical detection and visualization of generated text. *arXiv preprint arXiv:1906.04043*.

A. Gelman and Jennifer L. Hill. 2006. Data analysis using regression and multilevel/hierarchical models: Multilevel logistic regression. *Cambridge University Press,*.

Tatsunori B Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. *arXiv preprint arXiv:1904.02792*.

Roy S. Hessels, Chantal Kemner, Van Den Boomen Carlijn, and Ignace T. C. Hooge. 2016. The area-of-interest problem in eyetracking research: A noise-robust solution for face and sparse stimuli. *Behavior Research Methods*, 48(4):1694–1712.

Kenneth Holmqvist and Richard Andersson. 2017. *Eye-tracking: A comprehensive guide to methods, paradigms and measures*. Eye-tracking: A comprehensive guide to methods, paradigms and measures.

Ignace T. C. Hooge, Antje Nuthmann, Marcus Nystrm, Diederick C. Niehorster, Gijs A. Holleman, Richard Andersson, and Roy S. Hessels. 2025. The fundamentals of eye tracking part 2: From research question to operationalization. *Behavior Research Methods*, 57(2).

Ignace TC Hooge, Gijs A Holleman, Nina C Haukes, and Roy S Hessels. 2019. Gaze tracking accuracy in humans: One eye is sometimes better than two. *Behavior Research Methods*, 51(6):2712–2721.

Keren Gruteke Klein, Shachar Frenkel, Omer Shubi, and Yevgeni Berzak. 2025. Surprisal takes it all: Eye tracking based cognitive evaluation of text readability measures. *Preprint*, arXiv:2502.11150.

Minna. Lehtonen, Matti. Varjokallio, Henna. Kivikari, Annika.Hultén, Sami. Virpioja, Tero. Hakala, Mikko. Kurimo, Krista. Lagus, and Riitta. Salmelin. 2019. Statistical models of morphology predict eye-tracking measures during visual word recognition. *Memory & cognition*, 47(7):1245–1269.

R Likert. 1932. A technique for the measurement of attitudes. *archieves of psychology*, 22 140:1–55.

Chin-Yew Lin. 2004. Rouge: a package for automatic evaluation of summaries. In *Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004, Barcelona, Spain*, pages 74–81.

Angela Lopez-Cardona, Sebastian Idesis, Miguel Barreda-Ngeles, Sergi Abadal, and Ioannis Arapakis. 2025. Oasst-etc dataset: Alignment signals from eye-tracking analysis of llm responses.

Sandeep Mathias, Diptesh Kanojia, Kevin Patel, Samarth Agarwal, Abhijit Mishra, and Pushpak Bhattacharyya. 2018. Eyes are the windows to the soul: Predicting the rating of text quality using gaze behaviour. *Preprint*, arXiv:1810.04839.

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the uniform information density hypothesis. *arXiv preprint arXiv:2109.11635*.

MM Mukaka. 2012. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi medical journal : the journal of Medical Association of Malawi*, 24(3):69—71.

Marcus Nyström and Kenneth Holmqvist. 2010. An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data. *Behavior Research Methods*, 42(1):188–204.

Byung Doh Oh and William Schuler. 2022. Entropy- and distance-based predictors from gpt-2 attention patterns predict reading times over and above gpt-2 surprisal. *arXiv e-prints*.

OpenAI. 2023. Gpt-3.5 turbo. `https://platform.openai.com/docs/models/gpt-3-5`. Accessed: 2025-05-15.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc Meeting of the Association for Computational Linguistics*.

K. Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372–422.

K. Rayner. 2009. The 35th sir frederick bartlett lecture: Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology*, 62(8):1457–1506.

E D. Reichle, A. Pollatsek, D L. Fisher, and K. Rayner. 1998. Toward a model of eye movement control in reading. *Psychological review*, 105(1):125–57.

Gideon E. Schwarz. 1978. Estimating the dimension of a model. *The Annals of Statistics*, 6(2).

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. 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):e2307876121.

Heather Sheridan and Erik D. Reichle. 2016. An analysis of the time course of lexical processing during reading. *Cognitive Science*, 40(3):522–553.

Galit Shmueli. 2010. To explain or to predict? *Statistical science: A review journal of the Institute of Mathematical Statistics*.

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

Larry E. Toothaker. 1994. Multiple regression: Testing and interpreting interactions. *Journal of the Operational Research Society*, 45(1):119.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation (INLG)*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151.

Daniel Wiechmann, Yu Qiao, Elma Kerz, and Justus Mattern. 2022. Measuring the impact of (psycho-)linguistic and readability features and their spill over effects on the prediction of eye movement patterns. *Preprint*, arXiv:2203.08085.

Ethan Gotlieb Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2025. Testing the predictions of surprisal theory in 11 languages. *Preprint*, arXiv:2307.03667.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

A Details of the experiment

During the experiment, participants interact with a simplified interface to minimize distractions and accidental input errors. Custom keyboard bindings are used:

- **Navigation**: The `left` and `right` arrow keys move the focus between options.
- **Selection**: The `enter` key submits the chosen option.

After reading each text pair (displayed sequentially), participants select their preferred option using these keys. The task is self-paced, with no time limits or comprehension checks. The figure 3 shows the experimental interface and sequence.

B Score Bands and Grades from Various English Proficiency Tests

In summary, level B accounts for the largest proportion of 32.4%, followed by level C and D with 29.7% and 24.3% respectively, and the remaining level A and E accounts for 5.4% and 8.1%, a distribution that approximates a normal distribution. Here we present the score bands and grades Table 4 from various English proficiency tests for reference.

level	NCEE	CET-4	CET-6	IELTS	TOEFL
level A	/	/	/	8+	105+
level B	140+	600+	600+	7-7.5	90-105
level C	/	/	500-600	6-6.5	75-90
level D	130-140	500-600	425-500	5-5.5	60-75
level E	120-130	425-500	/	/	/

Table 4: Exam type and score levels for different exams. NCEE: National College Entrance Examination

Proficiency Level	Participant Proportion	Equivalent Test Scores
A (Advanced)	5.4%	IELTS 8+ / TOEFL 105+
B (Upper-Intermediate)	32.4%	IELTS 7–7.5 / TOEFL 90–105
C (Intermediate)	29.7%	IELTS 6–6.5 / TOEFL 75–90
D (Lower-Intermediate)	24.3%	IELTS 5–5.5 / TOEFL 60–75
E (Basic)	8.1%	CET-4 425–500

Table 5: English proficiency distribution and equivalent standardized test scores

C Scan path

Eye movement trajectories during reading were analyzed using the Spatial Temporal-DBSCAN clustering algorithm, which integrates temporal and spatial thresholds to cluster fixation points and filter noise. AOIs were defined as bounding boxes around individual words (with punctuation merged into preceding words). Mapping fixations to AOIs revealed:

D Metrics of dataset

D.1 Metrics over dataset

The figure 5 represents the distributions of attributes over models.

- **PDT Distribution**: As shown in the figure 5a, the distribution is unimodal, with a prominent peak at low PDT values and a long right tail. In other words, the majority of PDT values cluster in the lower range, and the percentage falls off sharply beyond the peak. The distribution is therefore skewed to the right, with the highest concentration at small PDT values.
- **BSF and Saccade Distribution**: As shown in the figure 5b, the BSF distribution is unimodal with a clear peak at a moderate frequency, and then drops off rapidly at higher frequencies. In contrast, the saccade frequency distribution peaks at a higher frequency but with a smaller maximum percentage, and it decays more gradually. In summary, BSF frequencies are mostly concentrated in the mid-range (producing a sharp peak), whereas saccade frequencies are more broadly spread with a flatter peak.
- **FT, Grading time and Reading time Distribution**: As shown in the figure 5c, the reading time distribution is unimodal, peaking around 19,000 ms and then gradually declining; it spans a wide range up to the highest time bins, indicating that some reading durations extend into tens of seconds. The fixation time distribution is relatively flat and appears to have two modest peaks: one near 15,000 ms (14%) and another around 21,000 ms (15%), suggesting a broad spread of fixation durations. In contrast, the grading time distribution is strongly peaked at very short durations: its peak is around 6,000 ms (about 33%) and it falls off steeply thereafter. This indicates that most grading has a short duration.

D.2 Metrics over model

The figure 6 represents the distributions of attributes over models. The left subfigure 6a illustrates the **Negative Log-Likelihood (NLL)** distribution across different language models, which

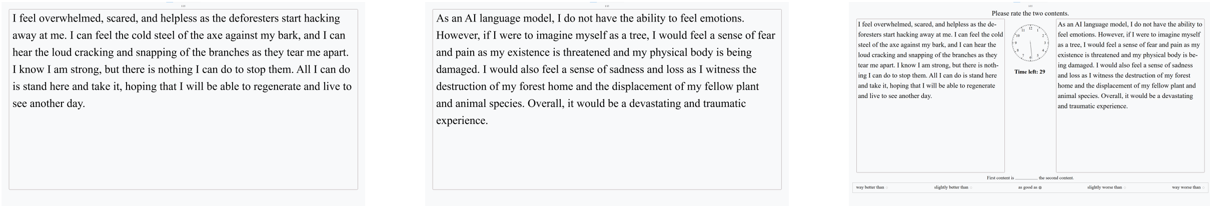


Figure 3: Experiment interface. Participants press `start` to begin reading a text pair, then `end` to proceed to rating. Keyboard bindings (`left`/`right` arrows and `enter`) simplify option selection.

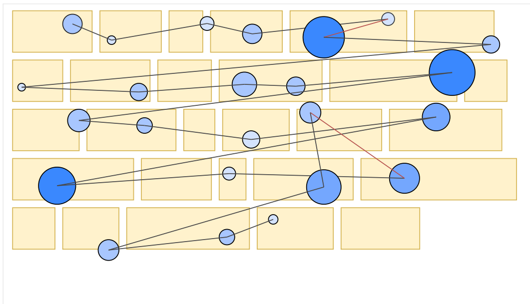


Figure 4: An example of scanning a path, where a box represents an Area-of-Interest (AOI). The circles represent the gaze locations, where a larger circle indicates a longer fixation time. The red lines represents regressions.

quantifies the uncertainty of text generation. Key observations include:

- **GPT-4's Predictability**: GPT-4 exhibits the narrowest NLL distribution, centered around a mean of 150 with a small standard deviation ($\sigma = 25$). This indicates that GPT-4 generates text with consistently high predictability, aligning with its reputation for producing coherent and contextually stable outputs.

- **Claude-v1's Variability**: Claude-v1 displays a skewed NLL distribution with a higher mean (220) and larger σ (80). The presence of frequent high-NLL values indicates that its generated text often contains unpredictable or less coherent segments. This variability may stem from Claude-v1's approach to generating more creative or diverse content, which can occasionally lead to linguistic discontinuities.

- **LLaMA-13B and Alpaca-13B**: These open-source models show broader NLL distributions compared to GPT-4 but are more concentrated than Claude-v1. The higher NLL values (relative to GPT-4) suggest greater lexical uncertainty, which may reflect their smaller training datasets or less refined fine-tuning.

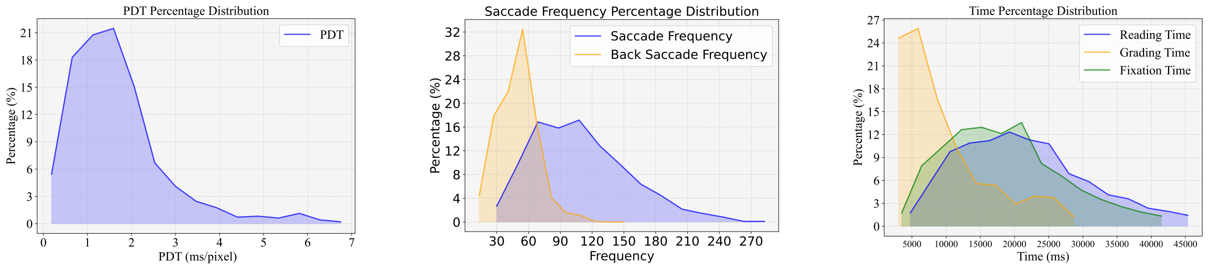
The right subfigure 6b compares the **MT-score scores (1.0–2.0)** with the experimental rating scores (EXP, 1.0–2.0), providing insights into human preference alignment across models:

- **GPT-3.5-Turbo's Dominance**: GPT-3.5-Turbo has the largest number of texts and a balanced distribution across MT-score scores, with a high proportion of MT-score=2.0 (50%) and EXP $=$ 2.0 (45%). This suggests strong alignment between automated MT-score scores and human judgments, likely due to its ability to generate fluent, task-relevant responses. The concentration of EXP scores at 1.75–2.0 highlights its popularity among participants, possibly driven by its optimal balance of readability (low NLL) and informativeness (moderate surprisal).

- **GPT-4's Uniform Quality**: GPT-4 has fewer texts but exhibits a uniform distribution of MT-score and EXP scores, with 50% rated MT-score=2.0 and no scores below MT-score=1.5. This reflects its consistent high quality, as human raters rarely deemed its outputs subpar.

- **LLaMA-13B's Lower Performance**: LLaMA-13B shows a dominance of low scores (MT-score=1.0: 60%), indicating poor human preference. This correlates with its higher NLL values, suggesting that less predictable text structure leads to increased cognitive effort (e.g., longer Fixation Time) and lower perceived quality.

- **Model-Specific Trends**: Vicuna-13B-v1.2 and Alpaca-13B show moderate performance, with MT-score=1.5 as their modal score. Their EXP distributions are slightly skewed toward higher values, suggesting that fine-tuning on instruction-following tasks improves readability compared to base models like LLaMA-13B.

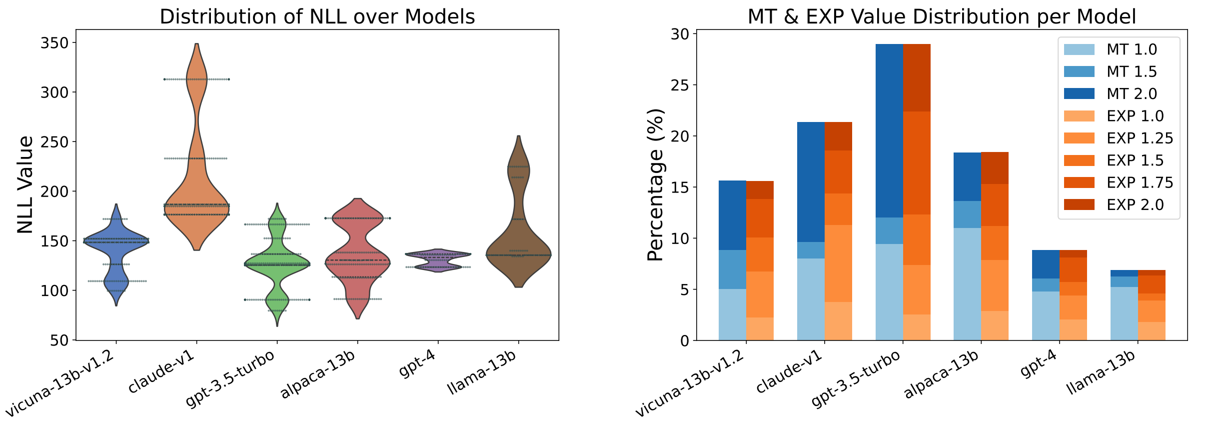


(a) The distribution of PDT over dataset.

(b) The distribution of BSF and saccade over dataset

(c) The distribution of FT, rating time and reading time over dataset.

Figure 5: The distributions of attributes over models.



(a) The distribution of nll over models.

(b) The distribution of MT-Bentch score (values of 1.0, 1.5, 2.0) and actual rating score on dataset (the exp values of 1.0, 1.25, 1.5, 1.75, 2.0).

Figure 6: The distributions of attributes over models.

Cross-Figure Insights

- **Correlation Between NLL and Human Ratings**: Models with lower NLL (e.g., GPT-4) generally receive higher EXP scores, supporting the hypothesis that token-level predictability contributes to perceived quality. However, the combination of NLL and eye-tracking metrics (e.g., PDT, BSF) in Model $M_{\mathrm{EYE+NLL}}$ (Table 3) demonstrates that gaze data adds unique explanatory power beyond linguistic features alone.

- **Implications for LLM Evaluation**: The figures underscore the value of incorporating eye-tracking into LLM assessment. For example, Claude-v1's high NLL variability may not be fully captured by traditional metrics, but eye-movement patterns (e.g., inconsistent fixation durations) can reveal hidden weaknesses in text flow.

By integrating quantitative NLL distributions with qualitative human ratings, these figures provide a comprehensive view of how model architecture influences both linguistic predictability and reader experience, reinforcing the necessity of multimodal evaluation frameworks like GREAT.