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

Xiaoshan He[1]*, Xiaoqun Liu[1]*, Haodong He[1]*, Yu Wang [2] Yang Xu [1]
{12110848,12110943,12111627,xuyang}@mail.sustech.edu.cn
y.wang@uni-bielefeld.de
[1]Department of Computer Science and Engineering,
Southern University of Science and Technology
[2]Digital Linguistics Lab, Department of Linguistics,
Bielefeld University

Abstract

Eye movement offers valuable insights into human visual attention during assessment of machine-generated texts, 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 features during screen reading of passages generated by large language models (LLMs). The dataset includes raw eye-movement recordings, reading-time measurements, and post-reading evaluations for LLM-generated passage pairs, alongside rigorous validation metrics. The collected eye-movement features 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-movement can act as an effective source of information that complements probabilistic metrics, for the task of automatic text quality assessment. The full dataset and some processing code are publicly available at https://github.com/qwurd231/GREAT.

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 technique in psycholinguistics for studying the underlying cognitive processes during reading activities. Eye-movement features 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, the direct applications of eye-tracking to the evaluation of machine-generated text—especially from modern LLMs—remain relatively underexplored.

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

In this study, we design a set of screen-reading experiments following the classical psycholinguistic paradigm, and focus on the instantaneous eye-movement reaction to the model-generated texts displayed, aiming to bridge eye-movement metrics with NLP evaluation. We introduce GREAT (Gaze Responses for Evaluating AI Texts), a dataset that captures the eye movements from human readers as they read and evaluate LLM-generated text responses. Based on the MT-Bench dataset, which provides human preference labels for pairs of LLM-generated texts. Our dataset includes not only behavioral data, such as reading passes and fixation

*Equal contribution

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 470–486

patterns, but also post-reading quality judgment scores. We systematically analyze the relations between the two components, aiming to uncover how eye movements measured in multiple dimensions—both temporal and spatial—can serve as proxies for human assessments of text quality.

Our central research question is: To what extent can eye-movement features predict the perceived quality of LLM-generated text? Especially compared to or combined with other features such as negative log-likelihood (NLL). To explore this, we evaluate the predictive power of several common eye-movement 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 main contributions:

- **A novel dataset (GREAT)** with fine-grained eye-movement features collected from human participants reading LLM-generated text pairs.
- **Validation experiments:** We demonstrate that eye-movement features significantly enhance the predictive power of text quality assessment when combined with common statistical metrics like NLL.

2 Related Works

Eye-tracking technology has been widely used to study text readability and comprehension. Prior studies show that eye-movement measures reflect cognitive load and processing difficulty when reading machine-translated text, supporting their value as evaluation signals (Colman et al., 2022). Large-scale resources such as EMTeC further extend this paradigm to natural reading of machine-generated texts (Bolliger et al., 2025). These studies demonstrate the value of gaze signals for evaluating machine-translated text. However, they mainly focus on translation quality, post-editing effort, or reception of translated documents, while our work studies open-ended LLM-generated responses and asks whether eye-movement features can predict human preference judgments over generated texts.

2.1 Eye-Movement Metrics

Two key metrics of eye-movement during reading are **fixations** and **backward saccades**. Fixations, during which the eyes pause to read, enable information uptake, while backward saccades (regressions) reflect rereading of earlier material. Prior research shows that as text difficulty increases, fixation durations lengthen, saccade lengths shorten, and backward saccades become more frequent (Rayner, 1998, 2009). These patterns are robust across languages and correlate with traditional readability scores (Baazeem et al., 2021, 2025; Atvars and Aigars, 2017), suggesting that eye-movement data reliably capture text difficulty. Consequently, **reading time** is also expected to correlate with readability, and is hence included as a variable in our analysis.

2.2 Human judgements of generated texts

Human evaluation is one of the major approaches to assessing model-generated texts (Celikyilmaz et al., 2020), but it is challenged by several limitations: the lack of preferences over diverse texts (Hashimoto et al., 2019), the need for multiple assessors (van der Lee et al., 2019), and biases arising from sample selection and sequential effects (van der Lee et al., 2021). The relationship between the negative log-likelihood (NLL) values of texts (or surprisal, entropy) and cognitive efforts required for reading, measured by reading time, has long been established (Smith and Levy, 2013). Recently, NLL has been utilized to develop cognitive-inspired evaluation frameworks of text readability together with eye-movement features (Klein et al., 2025). Similarly, EMTeC, a corpus of naturalistic eye-movements-while-reading corpus of human subjects reading machine-generated texts is collected (Bolliger et al., 2025). The main differences from our work and the existing studies are that we focus on the predictive effect of eye-movement features alone in modeling human perceived text quality, and the text data we use are annotated by relatively broad human judges.

3 Experiment Setup

3.1 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 QA dataset created for evaluating chatbots' conversational skills and instruction-following capabilities, comprising 30k machine-generated conversations with human preference annotations. We use MT-Bench under the default OpenAI API settings (no additional prompt engineering), where the system–user message template serves as the evaluation context. Each text pair corresponds to responses generated for the

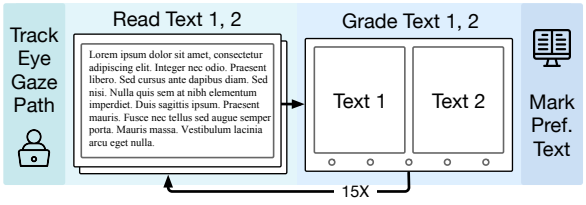


Figure 1: **Task workflow for dataset construction.** Participants completed 15 sessions, each involving the sequential reading of two texts while their eye gaze was recorded. After reading, participants rated their relative preference for the texts using a five-point scale displayed beneath the text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

same prompt in MT-Bench. We manually curated a subset of 39 texts to serve as reading materials. Each participant only read a subset of text pairs (15 pairs), and the assignment of texts to participants was randomized. As a result, each text was read by an average of 16.62 participants, in accordance with two guiding principles:

Text length To prevent scrolling or page flipping, texts were restricted to fit within a single screen, with an average length of 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 from sources such as Wikipedia and news articles. This ensures that the majority of subjects can understand the material without difficulty caused by domain knowledge.

3.2 Data Collection

3.2.1 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 where English is the medium of instruction (EMI), took part in the eye-tracking experiment. Beforehand, they completed a questionnaire on English proficiency; Appendix B summarizes the self-reported levels, compared with standard test score bands (e.g., TOEFL, IELTS, CET-4/6; see Table 5).

3.2.2 Apparatus

The experiment was conducted in a controlled environment: unrelated personnel were cleared, participants wore noise-cancelling headphones, and a 28 inches display (2560×1440 pixels resolution) was used. Before starting, participants completed eye-tracker calibration, and during the task they were instructed to remain still to ensure data quality. Eye movement data were recorded using Tobii Pro Spark eye tracker, operating at a sampling rate of 60 Hz.

3.2.3 Procedure

We recruited 38 participants for the eye-tracking experiment. After a demonstration video explaining the procedure, participants read texts at their own pace without comprehension testing, while asked to remain still. A six-point calibration was performed before the task to ensure data accuracy. Participants used custom keyboards to minimize distractions, prevent accidental touches, and reduce noise caused by looking for the keyboard. The experiment consisted of two stages: reading and rating. In the reading stage, participants used the keyboard to control the start and end of two readings. After reading the two texts, they rated the two texts on a five-point Likert scale (Likert, 1932), indicating their preference within 30 seconds to ensure attention. Each participant completed 15 cycles (15 text pairs, 30 texts in total, randomly displayed).

To ensure reliable preference selection, both texts were displayed simultaneously on the screen, and participants indicated their choice via on-screen options with a confirmation prompt. A one-second throttle measure was implemented to prevent accidental or repeated inputs. See more experimental details in the appendix A.

4 Data Processing

4.1 Data Cleaning

Denoising We remove three types of noise data: the samples with total reading durations beyond two standard deviations from the mean; samples with insufficient valid gaze points to control for technical artifacts; and cases where ratings were not provided in a timely manner. Applying these criteria removed 83 pairs, leaving 487 pairs of valid reading data for analysis. Our eye tracker records binocular gaze data, and we follow the established practice (Hooge et al., 2019) by averaging left and right eye signals to reduce systematic error. A common issue in eye-tracking experiments is *vertical drift*, i.e., a gradual displacement of recorded gaze coordinates over time (Carr et al., 2022; Chen et al., 2021; Frank and Aumeistere, 2024). We use a clustering method[1] to resolve this issue.

[1] `AgglomerativeClustering` from scikit-learn

Area of Interest (AOI) AOI is the basis for defining subsequent eye-movement features. We follow the conventions in literature and define AOI as the bounding box of a word (delimited by a space " "), with punctuation incorporated into the preceding word (Holmqvist et al., 2011; Hessels et al., 2016; Hooge et al., 2025). Specifically, gaze points outside the reading region (e.g., screen edges) are excluded, and remaining points are assigned to the nearest AOI.

Detecting microsaccades The first step of extracting saccade paths between words is to detect *microsaccades*. We employed a velocity-based algorithm (Engbert and Kliegl, 2003; Nyström and Holmqvist, 2010), where saccades are defined as velocity outliers relative to the overall distribution. To minimize noise and enhance detection stability, gaze data were first smoothed with a five-sample weighted moving average. Assuming an approximately normal velocity distribution, the detection threshold was set at two standard deviations above the mean. We find this a simple yet effective way to identify microsaccadic activities robustly.

Extracting saccade trajectories Based on the microsaccades detected, we extract the full saccade trajectories during reading, by adopting a time-space-based clustering method known as Spatial Temporal-DBSCAN (Birant and Kut, 2007), which clusters fixation points in both time and space using predefined thresholds while filtering out the noise. The resulting clusters are mapped to AOIs to construct the scan paths (Figure 4); further implementation details are provided in Appendix C.

4.2 Feature Construction

Based on the preprocessed data from the previous step, we construct a set of features that will be the candidate predictors for later modeling tasks:

Reading Time Total time for reading a complete text material on screen. The average reading time is 21.37 sec (SD = 9.35).

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) 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, we computed token-level surprisal (NLL) using the LLaMA-7B model (Touvron et al., 2023), as it provides a consistent estimate of text predictability across different passages (Hale, 2001; Levy, 2008). While Shannon entropy (Shannon, 1948) formally quantifies the expected uncertainty of the model's predictive distribution, NLL reflects the realized surprisal of the observed text. Empirically, GPT-4 texts exhibit the narrowest NLL distributions, suggesting stable confidence, whereas Claude-v1 (Anthropic, 2023) shows broader, skewed distributions, indicating higher variability.

MT-bench Score (MT-score) & Experiments score (EXP-score) 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 higher, the better, 1.5 means a tie, the finer scale allows us to distinguish weak from strong preferences, which is not possible with the original MT-Bench labels. The two scales are kept separate: EXP-score is the primary dependent variable in our models, while MT-score is used only for external validation (Section 4.3). The majority scores of MT demonstrate clear preferences between text pairs.

A similar pattern is observed in the EXP results. The gpt-3.5-turbo (OpenAI, 2023) contributes the largest number of texts, spanning the full range of MT and EXP scores, and particularly enriched in the cases (MT$=2.0$, EXP$=2.0$). 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.

4.3 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 with the standardized MT-Bench scores. After aligning the scoring scales—mapping EXP-scores: 1.0/1.25→1.0, 1.75/2.0→2.0, 1.5 unchanged—the matching accuracy reached 80%, comparable to Zheng et al. (Zheng et al., 2023). This high consistency supports the robustness and external validity of our experimental framework.

We further examined the link between fixation duration and lexical surprisal, 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 passes (Wilcox et al., 2023). To empirically evaluate this relationship, we computed Spearman correlation coefficients between surprisal values and fixation durations across the dataset (Mukaka, 2012), and observed weak but statistically significant positive associations ($\rho_{\text{Spearman}} = 0.072$, 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 suggests that additional factors—such as individual cognitive differences, reading strategies, task demands, and higher-level linguistic complexity (Sheridan and Reichle, 2016)—also play a role. Overall, the dataset thus reflects both psycholinguistically sound variance and the multifactorial nature of human reading behavior.

5 Experiments

Based on the dataset we have collected and cleaned, we further study how eye-movement features can predict human preferences in evaluating LLM-generated texts. Specifically, we use linear mixed-effects models as the basic approach (Section 5.3), and a neural reward model as an advanced attempt (Section 5.4).

5.1 Multicollinearity analysis and variable selection

The limited proportion of variance in the outcome measure explained by language model identity indicates that observed differences across models largely reflect architectural and training properties, rather than meaningful variance in our outcome measure. We therefore exclude model identity from subsequent analyses to better isolate the effects of linguistic and eye-movement features.

To prevent any potential multicollinearity issues in the modeling task of MT-score prediction, we compute Variance Inflation Factors (VIF) (Toothaker, 1994) to quantify the amount of variance each predictor is inflated with due to correlations with other predictors. We keep only those predictors with VIF ≤ 5, following common practice in multivariate data analysis (Murtagh and Heck, 2012; O'brien, 2007; James et al., 2013), ensuring independent and interpretable feature contributions.

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

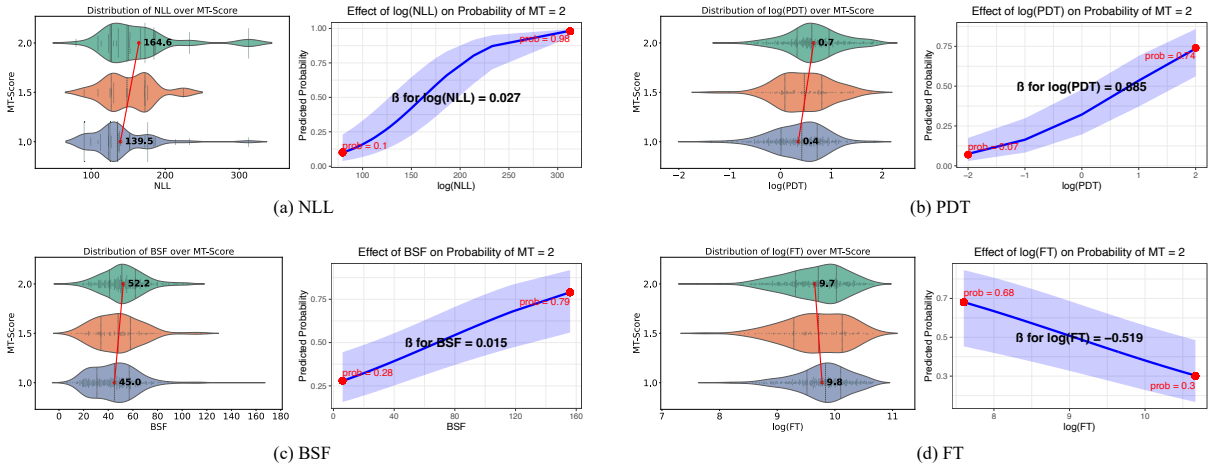
Table 1: The predictors with small VIF values.

As shown in Table 1, the final predictors we keep for the subsequent modeling tasks are: NLL (linguistic predictor) and three eye-movement metrics, PDT, BSF, and FT.

5.2 Hypotheses

We propose the following hypotheses regarding the relationship between eye-movement 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



(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).

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-movement metrics improves predictive accuracy beyond either metric individually, demonstrating their complementary contributions.

We posit that eyemovement-based features offer a richer and more direct window into readers' cognitive engagement with text than 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.

5.3 Mixed-effect linear models

This section presents mixed-effect regression models examining the performances of predicting human-judged text quality scores (MT-Score, defined in Section 4.2) using the four predictors from Table 1. Random intercepts are fit to account for variability across text generators and participants. The coefficients from single-predictor models are shown in Table 2. In summary, **all eye-movement features are significant predictors of text quality**, which strongly supports our hypotheses that directly measured cognitive efforts during language processing can reflect the high-level quality of input text. Details of each model are discussed as follows.

Model	β	SE	t-value	Effect Direction	p-value
M_{NLL}	2.699×10^{-2}	3.174×10^{-3}	8.504	Positive	1.833×10^{-17}
M_{PDT}	8.850×10^{-1}	1.467×10^{-1}	6.034	Positive	1.596×10^{-9}
M_{BSF}	1.520×10^{-2}	4.163×10^{-3}	3.651	Positive	2.612×10^{-4}
M_{FT}	-5.188×10^{-1}	1.477×10^{-1}	−3.514	Negative	4.418×10^{-4}

Table 2: Parameter estimates, standard errors, t-values, and 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}) We first set a baseline model M_{NLL}, which only uses token-level uncertainty, measured by NLL, on predicting text quality. The model is formulated 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 the variance in text generation capabilities of the different generator models from the MT-Bench dataset; ϵ is the residual error.

The coefficient for NLL is significantly positive, indicating that texts with higher token-level surprisal tend to receive higher MT-Scores. This find-

ing corroborates prior work suggesting that readers favor content that is more novel or informative (Gehrmann et al., 2019). M_{NLL} establishes a baseline for evaluating the subsequent models with eye-movement features added.

Model 1: Effect of Pixel Dwelling Time (M_{PDT}) PDT measures the average duration of visual attention for a reading session (Rayner, 1998). We fit the model M_{PDT} formulated as:

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

Different from M_{NLL}, this formula includes an additional random intercept for the human subject to account for the individual difference in text processing capabilities. A significant positive effect of PDT on MT-Score is found, indicating that longer visual attention (larger PDT value) is associated with higher subjective quality ratings (higher MT-Score). This finding supports hypothesis **H1**: the increased cognitive engagement, as reflected by PDT, reveals a more favorable human evaluation outcome.

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 more rereading behavior.

Model 3: Effect of Fixation Time (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:

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

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, as reflected by reduced fixation time, are perceived as having higher 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-movement 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-movement metrics—PDT, FT, and BSF—for brevity.

To assess the predictive capacity of different models for text quality grading, Table 3 reports results on four 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 (lower is better) (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.

Compared with the NLL-only baseline (M_{NLL}), all two-predictor combinations involving eye-movement metrics (e.g., M_{PDT+FT}, M_{FT+BSF}) demonstrate improved explanatory power across all evaluation metrics, confirming that these features provide further 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 parameter efficient model choice.

The full eye model (M_{EYE}: PDT+FT+BSF) shows a modest improvement over M_{PDT+FT}, though with a slightly higher BIC, indicating a minor trade-off in model parsimony. By combining the eye-movement metrics with NLL, the model $M_{EYE+NLL}$ achieves the best performance across all four evaluation criteria. The substantial increase in both goodness-of-fit measures suggests that eye-movement features serve as an important complement to NLL in predicting text quality grading, offering robust empirical support for **H4**.

Held-out Prediction To further validate our regression analysis, we evaluated predictive performance on held-out test sets (80/20 split, 5-fold cross-validation). Results show that combining NLL with eye-movement features improves correlation with human judgments (Spearman $\rho = 0.337$ vs. 0.308 with NLL alone) and explained variance ($R^2 = 0.125$ vs. 0.092). Full metrics are reported in Appendix F.

5.4 Reward-model-based prediction

Beyond the mixed-effects linear models, we also explore reward modeling–directly predicting human preferences (the judgement scores) by adding a regression head to a language model, a common practice in the reinforcement learning from human feedback (RLHF) (Christiano et al., 2017; Ouyang et al., 2022; Rafailov et al., 2023). Instead of fine-tuning the whole LLM, we only train a small MLP on top of frozen LLaMA embeddings.

Experimental setup We adopt a standard pairwise preference modeling framework, where the goal is to assign higher scores to the preferred text in each pair. The model is trained with a Bradley–Terry style pairwise loss (Bradley and Terry, 1952), defined as:

$$\mathcal{L} = -\log \sigma(r_i - r_j)$$

where r_i and r_j are the predicted scores for the preferred and non-preferred texts, respectively. Each text is represented by : (i) the final layer hidden states from the LLaMA model used for NLL computation, reduced from 4096-d to 256-d via PCA (Reimers and Gurevych, 2019). PCA both mitigates overfitting on our small dataset and removes redundant global variance (Arora et al., 2017; Mu et al., 2017), but may also discard fluency-related dimensions. To complement this, (ii) we explicitly include the NLL, which directly reflects linguistic predictability and fluency, unlike embeddings that primarily capture semantic information (Hale, 2001; Levy, 2008; Goodkind and Bicknell, 2018). (iii) We further incorporate optional features extracted from our eye-movement data - the same three metrics analyzed in (Section 5): Pixel Dwelling Time (PDT), Backward Saccade Frequency (BSF), and Fixation Time (FT), combined with text-based features through a gating layer (Srivastava et al., 2015; Arevalo et al., 2017). The model is evaluated with 5-fold cross-validation for robustness.

Main Result Table 4 shows the reward model's accuracy in predicting the EXP-score on the strong-preference subset (EXP $\in \{1, 2\}$) of our data, which constitutes about 30% of our dataset. We can see that adding eye-movement features consistently improves over the text-only baseline (0.601), with all feature combinations yielding positive gains. The best configuration (PDT+FT) achieves the highest Acc. (0.639, $p = 0.046$), while the joint (PDT+FT+BSF) setting also shows a significant gain ($p = 0.039$), based on bootstrap resampling (Efron and Tibshirani, 1994). Additional experiments with weak preferences, ties, and a simpler regression learning objective (Appendix E) show smaller but consistent gains, confirming the robustness of the effect.

Model Variant	Exp-Acc	p-value
$\text{Baseline}_{\text{embedding+NLL}}$	0.601±0.022	\
$+\text{EYE}_{\text{FT+BSF}}$	0.632±0.039	0.102
$+\text{EYE}_{\text{PDT+BSF}}$	0.633±0.033	0.057
$+\text{EYE}_{\text{PDT+FT+BSF}}$	0.637±0.012	0.039
$+\text{EYE}_{\text{PDT+FT}}$	$\mathbf{0.639 \pm 0.034}$	0.046

Table 4: Exp-Acc (mean ± std) on strong-preference samples, evaluated by **5-fold cross-validation repeated four times** with different random seeds. p-values are computed via 1,000-sample bootstrap tests against the text-only baseline.

5.5 Correlations with other variables

The variable `model_name` (e.g., Claude-v1, GPT-3.5-Turbo, GPT-4, LLaMA-13B, Vicuna-13B-v1.2 (Chiang et al., 2023), and Alpaca-13B) shows strong correlation with MT-scores, and linear models including it achieve higher R-squared values. However, since `model_name` reflects system identity rather than a mechanistic predictor of text quality, its inclusion would confound interpretability by attributing variance to label information rather than cognitive processing (Gelman and Hill, 2006; Shmueli, 2010). We therefore exclude `model_name` from the main predictive models to control for model generation capabilities and focusing on cognitive processing indicators.

6 Conclusions

In this study, we introduced the GREAT dataset, an eye-movement annotated corpus of human subjects reading LLM-generated texts. The dataset comprises a comprehensive set of eye-movement features collected from controlled screen reading experiments. Our preliminary analysis of the data demonstrates that eye-movement features, such as reading time, forward/backward saccades, and fixation time, are significant predictors of human judgments of text quality. Among these features, the effect of Pixel Dwelling Time (PDT) is the most significant – almost the same level as that of the NLL (surprisal) of texts. The GREAT dataset offers new insights into how readers interact with and evaluate machine-generated content, highlighting the potential of including behavioral data such as eye-movement into the loop of NLG evaluation.

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 in the strict sense, 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 rat-

ings 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.

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. 2017. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *International conference on learning representations*.

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 S Bolliger, Patrick Haller, Isabelle CR Cretton, David R Reich, Tannon Kew, and Lena A Jäger. 2025. Emtec: A corpus of eye movements on machine-generated texts. *Behavior Research Methods*, 57(7):189.

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

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

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.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Toon Colman, Margot Fonteyne, Joke Daems, Nicolas Dirix, and Lieve Macken. 2022. Geco-mt: The ghent eye-tracking corpus of machine translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 29–38.

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.

Bradley Efron and Robert J Tibshirani. 1994. *An introduction to the bootstrap*. Chapman and Hall/CRC.

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. In *Proceedings of the 57th annual meeting of the association for computational linguistics: system demonstrations*, pages 111–116.

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

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 (CMCL 2018)*, pages 10–18.

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*.

Tatsunori B Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1689–1701.

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, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost Van de Weijer. 2011. *Eye tracking: A comprehensive guide to methods and measures*. oup Oxford.

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.

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. 2013. *An introduction to statistical learning: with applications in R*, volume 103. Springer.

Keren Gruteke Klein, Shachar Frenkel, Omer Shubi, and Yevgeni Berzak. 2025. Eye tracking based cognitive evaluation of automatic readability assessment measures. In *Computational Psycholinguistics Meeting 2025*.

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.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

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-Ángeles, Sergi Abadal, and Ioannis Arapakis. 2025. Oasst-etc dataset: alignment signals from eye-tracking analysis of llm responses. *Proceedings of the ACM on Human-Computer Interaction*, 9(3):1–29.

Sandeep Mathias, Diptesh Kanojia, Kevin Patel, Samarth Agrawal, Abhijit Mishra, and Pushpak Bhattacharyya. 2018. Eyes are the windows to the soul: Predicting the rating of text quality using gaze behaviour. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2352–2362, Melbourne, Australia. Association for Computational Linguistics.

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*, pages 963–980.

Jiaqi Mu, Suma Bhat, and Pramod Viswanath. 2017. All-but-the-top: Simple and effective postprocessing for word representations. *arXiv preprint arXiv:1702.01417*.

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.

Fionn Murtagh and André Heck. 2012. *Multivariate data analysis*. Springer Science & Business Media.

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. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9324–9334.

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

OpenAI et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Robert M O'brien. 2007. A caution regarding rules of thumb for variance inflation factors. *Quality & quantity*, 41(5):673–690.

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*.

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. *Advances in neural information processing systems*, 36:53728–53741.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

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.

Claude E Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

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.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Highway networks. *arXiv preprint arXiv:1505.00387*.

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 G Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

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 calibration phase, calibration points appeared at multiple locations on the display surface, aligned with the stimulus plane using default settings. To ensure calibration accuracy, we applied participant-specific adjustments, such as modifying chair and screen height to align the participant's head with the screen center, supplemented by the device's built-in mechanisms. The calibration process included a data collection phase followed by an optimization phase, during which the system calculated deviations between recorded gaze data and actual calibration targets, adjusting parameters in its internal 3D eye model. A pass/fail test served as an indicator of calibration quality, which is inspected by us every time in the display panel, though not recorded. 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.

During reading tasks, participants were instructed to remain still and maintain their posture from calibration. Additionally, the chosen text configuration with a large font size (80px) , font-family(Times New Roman) and line spacing (40px) defined broad Areas of Interest (AOIs), which helped tolerate small gaze deviations due to minor movements. 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 5 from various English proficiency tests for reference.

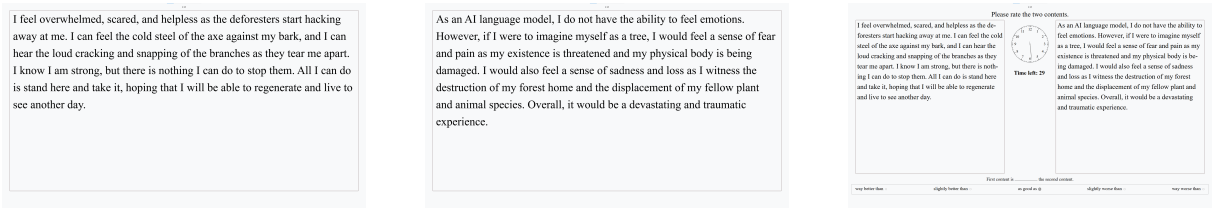


Figure 3: Experiment interface. Participants press `start` to begin reading a text pair, then `end` to proceed to rating. Keyboard bindings (`left`/`right` arrows and `enter`) simplify option selection.

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 5: 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 6: 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:

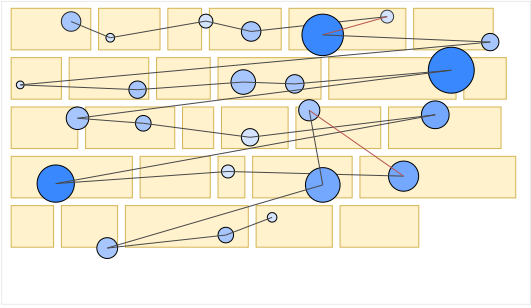


Figure 4: An example of scanning a path, where a box represents an Area-of-Interest (AOI). The circles represent the gaze locations, where a larger circle indicates a longer fixation time. The red lines represents regressions.

D Metrics of dataset

D.1 Metrics over dataset

Although our experimental (EXP) labels are collected on a subset of MT-bench items, the scoring scales differ. MT-bench provides absolute quality ratings based on a fixed rubric, whereas our EXP scores are relative preference judgments (with values in 1,1.25,1.5,1.75,2) designed to capture both strong and weak preferences. Thus, EXP and MT scores are not directly comparable in magnitude, but complementary in evaluation perspective. The figure 5 represents the distributions of attributes over dataset.

- **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 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.

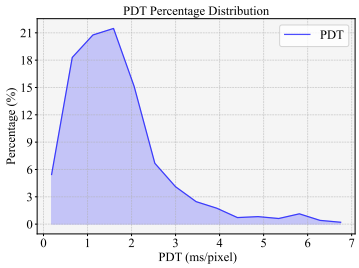
The right subfigure 6b compares the **MT-score scores (1.0–2.0)** with the experimental rating scores (EXP, 1.0–2.0), providing insights into human preference alignment across models:

- **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.

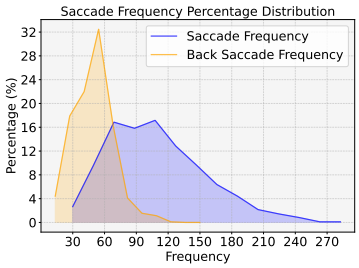
Cross-Figure Insights

- **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 $M_{\text{EYE+NLL}}$ (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.

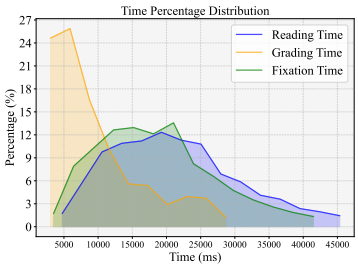
By integrating quantitative NLL distributions with qualitative human ratings, these figures provide a comprehensive view of how model architecture influences both linguistic predictability and



(a) The distribution of PDT over dataset.

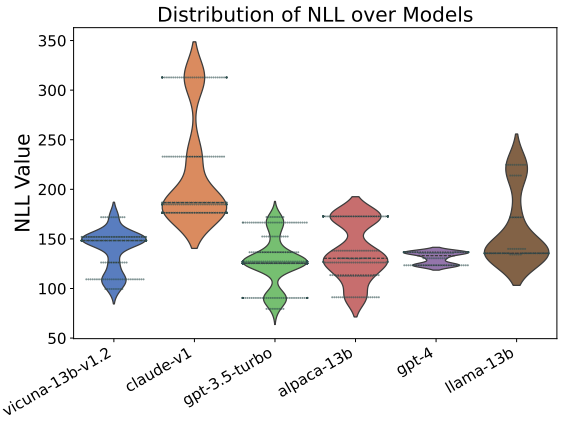


(b) The distribution of BSF and saccade over dataset

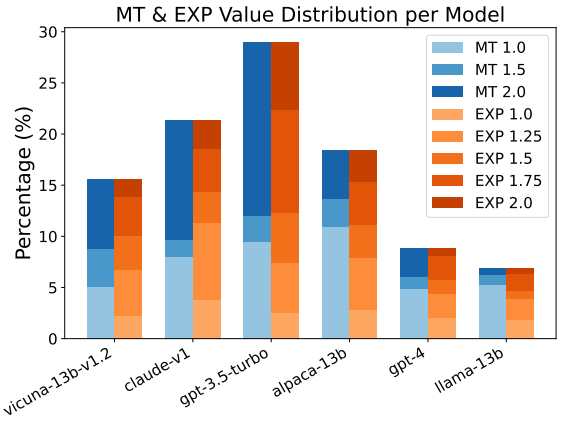


(c) The distribution of FT, rating time and reading time over dataset.

Figure 5: The distributions of attributes over models.



(a) The distribution of nll over models.



(b) The distribution of MT-Bench 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.

reader experience, reinforcing the necessity of multimodal evaluation frameworks like GREAT.

E Reward Model Additional Experiment

Table 7 reports reward-model performance when training on all samples, including weak preferences (exp = 1.25 or 1.75) and ties (exp = 1.5). We compare two training objectives: (i) pairwise loss and (ii) regression with continuous labels.

For pairwise training, we extend the standard Bradley–Terry loss to handle tie cases. Specifically, for non-tie pairs, we use the standard objective

$$\mathcal{L}_{\text{pair}} = -\log \sigma(r_i - r_j)$$

where r_i, r_j are the predicted scores for the two texts and σ is the sigmoid function. For ties (exp $=$ 1.5), we adopt a soft constraint that encourages the two predictions to be close:

$$\mathcal{L}_{\text{tie}} = \alpha \, |r_i - r_j|$$

with $\alpha = 0.5$. The final hybrid loss is:

$$\mathcal{L} = (1 - \mathbf{1}_{\text{tie}}) \cdot \mathcal{L}_{\text{pair}} + \mathbf{1}_{\text{tie}} \cdot \mathcal{L}_{\text{tie}}$$

where $\mathbf{1}_{\text{tie}} = 1$ if the pair is labeled as a tie, and 0 otherwise. This design treats ties as a soft constraint, preventing the model from enforcing an arbitrary ordering while still utilizing these samples.

We report three metrics: (a) **Exp-Acc**, overall agreement with experimental preference scores; (b) **Strong-Acc**, restricted to strong-preference pairs (exp = 1 or 2); and (c) **Weak-Acc**, restricted to weak-preference pairs (exp = 1.25 ,1.5, 1.75). Weak-Acc is particularly challenging, since these labels reflect subtle and less consistent preferences. For both strong and weak subsets, accuracy is computed based on the direction of the predicted preference (i.e., which text is rated higher), while ties (exp = 1.5) are treated as correct if the predicted score difference falls within a small margin.

Across both objectives, gaze features yield small but consistent gains $\approx$ +0.58–1.2 points) over the text-only baseline. In the **pairwise setup**, the combined gaze feature set (PDT+BSF) achieves the best overall performance accuracy

(Exp-Acc $= 0.517$ vs. 0.505 baseline) , while the combined gaze feature set (PDT+FT+BSF) achieves the best strong-preference performance accuracy (StrongAcc $= 0.654$ vs. 0.622 baseline). A bootstrap significance test (1,000 iterations) confirms this improvement to be statistically reliable ($p = 0.011$, 95% CI: [0.004, 0.060]), indicating that gaze features provide complementary cues to textual signals even under noisy supervision.

For the **regression objective**, the BSF+FT gaze combination yields the largest gain on overall pairs (Exp-Acc $= 0.497$ vs. 0.480 baseline), while the PDT+FT gaze combination yields the largest gain on weak-preference pairs (WeakAcc $= 0.437$ vs. 0.414 baseline), also statistically significant ($p = 0.024$, 95% CI: [0.000, 0.047]). This suggests that gaze information is especially useful for modeling subtle or uncertain human judgments, where text-only representations provide limited discrimination.

This suggests that regression can leverage continuous labels when judgments are uncertain, whereas pairwise loss is more stable for clear preference cases. Although overall improvements are smaller than in strong-preference cases, the consistent trend across feature combinations demonstrates that gaze signals are robust even under noisier supervision.

Model Variant	Exp-Acc	Strong-Acc	Weak-Acc
Pairwise			
$\text{Baseline}_{\text{embedding+NLL}}$	0.505±0.005	0.622±0.014	0.451±0.001
$+\text{Gaze}_{\text{PDT+FT+BSF}}$	0.515±0.023	0.654±0.015	0.454±0.028
$+\text{Gaze}_{\text{PDT+FT}}$	0.513±0.016	0.632±0.007	0.459±0.023
$+\text{Gaze}_{\text{FT+BSF}}$	0.516±0.005	0.639±0.007	0.461±0.006
$+\text{Gaze}_{\text{PDT+BSF}}$	0.517±0.009	0.641±0.008	0.462±0.012
Regression			
$\text{Baseline}_{\text{embedding+NLL}}$	0.480±0.007	0.628±0.021	0.414±0.004
$+\text{Gaze}_{\text{PDT+FT+BSF}}$	0.483±0.004	0.615±0.010	0.425±0.004
$+\text{Gaze}_{\text{PDT+FT}}$	0.496±0.010	0.620±0.006	0.437±0.015
$+\text{Gaze}_{\text{FT+BSF}}$	0.497±0.013	0.649±0.021	0.429±0.013
$+\text{Gaze}_{\text{PDT+BSF}}$	0.492±0.007	0.643±0.031	0.424±0.014

Table 7: Reward-model performance trained on all samples, including weak preferences and ties. Each model is evaluated using **5-fold cross-validation repeated three times** with different random seeds, and mean ± standard deviation are reported. Exp-Acc measures overall agreement with experimental preference scores, Strong-Acc corresponds to strong-preference pairs (exp $\in \{1, 2\}$), and Weak-Acc covers weak-preference and tie pairs (exp $\in \{1.25, 1.5, 1.75\}$). Gaze features provide consistent, statistically significant improvements, particularly in strong and weak subsets.

F Held-out

To complement the mixed-effects regression analysis, we conducted a held-out evaluation using an 80/20 split with 5-fold cross-validation. Models were trained with NLL, eye-movement features (EYE), and their combination (NLL+EYE). Table 8 reports results across multiple metrics (RMSE, MSE, Spearman correlation, R^2). We observe that: NLL alone provides a modest predictive signal, consistent with surprisal-based accounts of reading difficulty. Eye-movement features alone achieve comparable performance, indicating that gaze carries independent information about text quality. The combination (NLL+EYE) yields the best performance across all metrics, with the largest gain in Spearman correlation (0.337) and R^2 (0.125). Although the improvements are moderate, they consistently demonstrate that gaze provides complementary predictive value beyond NLL.

Model	CV RMSE	Spearman Correlation	R^2	MSE
NLL	0.455	0.308	0.092	0.199
EYE	0445	0.229	0.082	0.202
NLL+EYE	0.442	**0.337**	**0.125**	**0.192**

Table 8: **Held-out predictive performance** (80/20 split, 5-fold CV). Combining eye-movement features with NLL improves correlation and explained variance.