# Happy Are Those Who Grade without Seeing: A Multi-Task Learning Approach to Grade Essays Using Gaze Behaviour

**Sandeep Mathias**[♠], **Rudra Murthy**[♠,◇], **Diptesh Kanojia**[♠,♣], **Abhijit Mishra**[◇], **Pushpak Bhattacharyya**[♠]
[♠] Department of Computer Science, Indian Institute of Technology, Bombay
[◇] IBM Research, India
[♣] IITB-Monash Research Academy
{sam,rudra,diptesh,pb}@cse.iitb.ac.in, abhijitmishra.530@gmail.com

## Abstract

The gaze behaviour of a reader is helpful in solving several NLP tasks such as automatic essay grading. However, collecting gaze behaviour from readers is costly in terms of time and money. In this paper, we propose a way to improve automatic essay grading using gaze behaviour, which is learnt at run time using a multi-task learning framework. To demonstrate the efficacy of this multi-task learning based approach to automatic essay grading, we collect gaze behaviour for 48 essays across 4 essay sets, and learn gaze behaviour for the rest of the essays, numbering over 7000 essays. Using the learnt gaze behaviour, we can achieve a statistically significant improvement in performance over the state-of-the-art system for the essay sets where we have gaze data. We also achieve a statistically significant improvement for 4 other essay sets, numbering about 6000 essays, where we have no gaze behaviour data available. Our approach establishes that learning gaze behaviour improves automatic essay grading.

## 1 Introduction

Collecting a reader's psychological input can be very beneficial to a number of Natural Language Processing (NLP) tasks, like complexity (Mishra et al., 2017; González-Garduño and Søgaard, 2017), sentence simplification (Klerke et al., 2016), text understanding (Mishra et al., 2016), text quality (Mathias et al., 2018), parsing (Hale et al., 2018), *etc.* This psychological information can be extracted using devices like eye-trackers, and electroencephalogram (EEG) machines. However, one of the challenges in using reader's information involves collecting the psycholinguistic data itself.

In this paper, we choose the task of automatic essay grading and show how we can predict the score that a human rater would give using both text and ***learnt*** gaze behaviour. An essay is a piece of text, written in response to a topic, called a prompt. Automatic essay grading is assigning a score to the essay using a machine. An essay set is a set of essays written in response to the same prompt.

Multi-task learning (Caruana, 1998) is a machine learning paradigm where we utilize auxiliary tasks to aid in solving a primary task. This is done by exploiting similarities between the primary task and the auxiliary tasks. **Scoring the essay** is the *primary task* and **learning gaze behaviour** is the *auxiliary task*.

Using gaze behaviour for a very small number of essays **(less than 0.7% of the essays in an essay set)**, we see an improvement in predicting the overall score of the essays. We also use our gaze behaviour dataset to run experiments on **unseen** essay sets - *i.e.,* essay sets which have **no gaze behaviour data** - and observe improvements in the system's performance in automatically grading essays.

**Contributions** The main contribution of our paper is describing how we use gaze behaviour information, in a multi-task learning framework, to automatically score essays outperforming the state-of-the-art systems. We will also release the gaze behaviour dataset[1] and code[2] - the first of its kind, for automatic essay grading - to facilitate further research in using gaze behaviour for automatic essay grading and other similar NLP tasks.

### 1.1 Gaze Behaviour Terminology

An ***Interest Area*** (IA) is an area of the screen that we are interested in. These areas are where some text is displayed, and not the white background on the left/right, as well as above/below the text. **Each word** is a separate and unique IA.

[1] Gaze behaviour dataset: http://www.cfilt.iitb.ac.in/cognitive-nlp/
Essays: https://www.kaggle.com/c/asap-aes
[2] https://github.com/lwsam/ASAP-Gaze

A ***Fixation*** is an event when the reader's eye is focused on a part of the screen. For our experiments, we are concerned only with fixations that occur within the interest areas. Fixations that occur in the background are ignored.

A ***Saccade*** is the path of the eye movement, as it goes from one fixation to the next. There are two types of saccades - Progressions and Regressions. ***Progressions*** are saccades where the reader moves from the current interest area to a *later* one. ***Regressions*** are saccades where the reader moves from the current interest area to an *earlier* one.

The rest of the paper is organized as follows. Section 2 describes our motivation for using eye-tracking and learning gaze behaviour from readers, over unseen texts. Section 3 describes some of the related work in the area of automatic essay grading, eye tracking and multi-task learning. Section 4 describes the gaze behaviour attributes used in our experiments, and the intuition behind them. We describe our dataset creation and experiment setup in Section 5. In Section 6, we report our results and present a detailed analysis. We present our conclusions and discuss possible future work in Section 7.

## 2 Motivation

Mishra and Bhattacharyya (2018), for instance, describe a lot of research in solving multiple problems in NLP using gaze behaviour of readers. **However**, most of their work involves collecting the gaze behaviour data first, and then splitting the data into training and testing data, before performing their experiments. While their work did show significant improvements over baseline approaches, across multiple NLP tasks, collecting the gaze behaviour data would be quite expensive, both in terms of time and money.

Therefore, we ask ourselves: "***Can we learn gaze behaviour, using a small amount of seed data, to help solve an NLP task?***" In order to use gaze behaviour on a large scale, we need to be able to *learn* it, since we can not ask a user to read texts every time we wish to use gaze behaviour data. Mathias et al. (2018) describe using gaze behaviour to predict how a reader would rate a piece of text (which is similar to our chosen application). Since they showed that gaze behaviour can help in predicting text quality, we use multi-task learning to simultaneously learn gaze behaviour information (auxiliary task) as well as score the essay (the primary task). However, they **collect all their gaze behaviour data *a priori***, while ***we try to learn the gaze behaviour of a reader*** and use what we learn from our system, for grading the essays. Hence, while they showed that gaze behaviour *could* help in predicting how a reader would score a text, their approach requires a reader to **read the text**, while our approach does not do so, ***during testing / deployment***.

## 3 Related Work

### 3.1 Automatic Essay Grading (AEG)

The very first AEG system was proposed by Page (1966). Since then, there have been a lot of other AEG systems (see Shermis and Burstein (2013) for more details). In 2012, the Hewlett Foundation released a dataset called the Automatic Student Assessment Prize (ASAP) AEG dataset. The dataset contains about 13,000 essays across eight different essay sets. We discuss more about that dataset later.

With the availability of a large dataset, there has been a lot of research, especially using neural networks, in automatically grading essays - like using Long Short Term Memory (LSTM) Networks (Taghipour and Ng, 2016; Tay et al., 2018), Convolutional Neural Networks (CNNs) (Dong and Zhang, 2016), or both (Dong et al., 2017). Zhang and Litman (2018) improve on the results of Dong et al. (2017) using co-attention between the source article and the essay for one of the types of essay sets.

### 3.2 Eye-Tracking

Capturing the gaze behaviour of readers has been found to be quite useful in improving the performance of NLP tasks (Mishra and Bhattacharyya, 2018). The main idea behind using gaze behaviour is the eye-mind hypothesis (Just and Carpenter, 1980), which states that whatever text the eye reads, that is what the mind processes. This hypothesis has led to a large body of work in psycholinguistic research that shows a relationship between text processing and gaze behaviour. Mishra and Bhattacharyya (2018) also describe some of the ways that eye-tracking can be used for multiple NLP tasks like translation complexity, sentiment analysis, etc.

Research has been done on using gaze behaviour at run time to solve downstream NLP tasks like sentence simplification (Klerke et al., 2016), readability (González-Garduño and Søgaard, 2018; Singh

et al., 2016), part-of-speech tagging (Barrett et al., 2016), sentiment analysis (Mishra et al., 2018; Barrett et al., 2018; Long et al., 2019), grammatical error detection (Barrett et al., 2018), hate speech detection (Barrett et al., 2018) and named entity recognition (Hollenstein and Zhang, 2019).

Different strategies have been adopted to alleviate the need for gaze behaviour at run time. Barrett et al. (2016) use token level averages of gaze features at run time from the Dundee Corpus (Kennedy et al., 2003), to alleviate the need for gaze behaviour at run time. Singh et al. (2016) and Long et al. (2019) predict gaze behaviour at the token-level prior to using it at run time. Mishra et al. (2018), González-Garduño and Søgaard (2018), Barrett et al. (2018), and Klerke et al. (2016), use multi-task learning to learn gaze behaviour along with solving the primary NLP task.

## 4 Gaze Behaviour Attributes

In our experiments, we use only a subset of gaze behaviour attributes described by Mathias et al. (2018) because most of the other attributes (like Second Fixation Duration[3]) were mostly 0, for most of the interest areas, and learning over them would not have yielded any meaningful results.

**Fixation Based Attributes** In our experiments, we use the Dwell Time (DT) and First Fixation Duration (FFD) as xation-based gaze behaviour attributes. Dwell Time is the total amount of time a user spends focusing on an interest area. First Fixation Duration is amount of time that a reader initially focuses on an interest area. Larger values for xation durations (for both DT and FFD) usually indicate that a word could be wrong (either a spelling mistake or grammar error). Errors would force a reader to pause, as they try to understand why the error was made (For example, if the writer wrote “shortcat” instead of “shortcut”.

**Saccade Based Attribute** In addition to the Fixation based attributes, we also look at a regression-based attribute - Is Regression (IR). This attribute is used to check whether or not a regression occurred from a given interest area. We don't focus on progression-based attributes, because the usual direction of reading is progressions. We are mainly concerned with regressions because they often occur when there is a mistake, or a need for disambiguation (like trying to resolve the antecedent of an anaphora).

**Interest Area Based Attributes** Lastly, we also use IA-based attributes, such as Run Count (RC) and if the IA was Skipped (Skip). The Run Count is the number of times a particular IA was xated on, and Skip is whether or not the IA was skipped. A well-written text would be read more easily, meaning a lower RC, and higher Skip (Mathias et al., 2018).

## 5 Dataset and Experiment Setup

### 5.1 Essay Dataset Details

We perform our experiments on the ASAP AEG dataset. The dataset has approximately 13,000 essays, across 8 essay sets. Table 1 reports the statistics of the dataset in terms of Number of Essays, Score Range, and Mean Word Count. The rst 4 rows in Table 1 are source-dependent response (SDR) essay sets, which we use to collect our gaze behaviour data. The other essays are used as unseen essay sets. SDRs are essays written in response to a question about a source article. For example, one of the essay sets that we use is based on an article called The Mooring Mast by Marcia Amidon Lüsted[4].

### 5.2 Evaluation Metric

| Essay Set | Number of Essays | Score Range | Mean Word Count |
|---|---|---|---|
| Prompt 3 | 1726 | 0-3 | 150 |
| Prompt 4 | 1770 | 0-3 | 150 |
| Prompt 5 | 1805 | 0-4 | 150 |
| Prompt 6 | 1800 | 0-4 | 150 |
| Prompt 1 | 1783 | 2-12 | 350 |
| Prompt 2 | 1800 | 1-6 | 350 |
| Prompt 7 | 1569 | 0-30 | 250 |
| Prompt 8 | 723 | 0-60 | 650 |
| Total | 12976 | 0-60 | 250 |

Table 1: Statistics of the 8 essay sets from the ASAP AEG dataset. We collect gaze behaviour data only for Prompts 3 - 6, as explained in Section 5.3. The other 4 prompts comprise our unseen essay sets.

For measuring our system's performance, we use Cohen's Kappa with quadratic weights - Quadratic Weighted Kappa (QWK) (Cohen, 1968) for the following reasons. Firstly, irrespective of whether we

[3] The duration of the xation when the reader xates on an interest area for the second time.

[4] The prompt is “Based on the excerpt, describe the obstacles the builders of the Empire State Building faced in attempting to allow dirigibles to dock there. Support your answer with relevant and speci c information from the excerpt.” The original article is present in Appendix A.

use regression, or ordinal classification. The final scores that are predicted by the system should be discrete scores. Hence, using Pearson Correlation would not be appropriate for our system. Secondly, F-Score and accuracy do not consider chance agreements unlike Cohen's Kappa. If we were to give everyone an average grade, we would get a positive value for accuracy and F-Score, but a Kappa value of 0. Thirdly, weighted Kappa takes into account the fact that the classes are ordered, i.e. $0 < 1 < 2 < \ldots$. Using unweighted Kappa would penalize a 0 graded as 4, as much as a 1. We use quadratic weights, as opposed to linear weights, because quadratic weights reward agreements and penalize mismatches more than linear weights.

### 5.3 Creation of the Gaze Behaviour Dataset

In this subsection, we describe how we created our gaze behaviour dataset, how we chose our essays for eye-tracking, and how they were annotated.

#### 5.3.1 Details of Texts

| Essay Set | 0 | 1 | 2 | 3 | 4 | Total |
|---|---|---|---|---|---|---|
| Prompt 3 | 2 | 4 | 5 | 1 | N/A | 12 |
| Prompt 4 | 2 | 3 | 4 | 3 | N/A | 12 |
| Prompt 5 | 2 | 1 | 3 | 5 | 1 | 12 |
| Prompt 6 | 2 | 2 | 3 | 4 | 1 | 12 |
| Total | 8 | 10 | 15 | 13 | 2 | 48 |

Table 2: Number of essays for each essay set which we collected gaze behaviour, scored between 0 to 3 (or 4).

As mentioned earlier in Section 5, we used only essays corresponding to prompts 3 to 6 of the ASAP AEG dataset. From each of the four essay sets, we selected 12 essays with a diverse vocabulary as well as all possible scores.

We use a greedy algorithm to select essays. For each essay set, we pick 12 essays, covering all score points with maximum number of unique tokens, as well as being under 250 words. Table 2 reports the distribution of essays with each score for each of the 4 essay sets that we use to create our gaze behaviour dataset.

To display the essay text on the screen, we use a large font size, so that (a) the text is clear, and (b) the reader's gaze is captured on the words which they are currently reading. Although, this ensures the clarity in reading and recording the gaze pattern in a more accurate manner, it also imposes a limitation on the size of the essay which can be used for our experiment. This is why, the longest essay in our gaze behaviour dataset is about 250 words. The original essays have their named entities anonymized. Hence, before running the experiments, we replaced the required named entities with placeholders (Eg. @NAME1 "Alex Smith", @PLACE1 "New Jersey", @MONTH1 "May", etc.)[5].

#### 5.3.2 Annotator Details

We used a total of 8 annotators, aged between 18 and 31, with an average age of 25 years. All of them were either in college, or had completed a Bachelor's degree. All but one of them also had experience as a teaching assistant. The annotators were fluent in English, and about half of them had participated earlier, in similar experiments. The annotators were adequately compensated for their work[6].

To assess the quality of the individual annotators, we evaluated the scores they provided against the ground truth scores, i.e., the scores given by the original annotators. The QWK measures the agreement between the annotators and the ground truth score. Close is the number of times (out of 48) in which the annotators either agreed with the ground truth scores, or differed from them by at most 1 score point. Correct is the number of times (out of 48) in which the annotators agreed with the ground truth scores. The mean values for the 3 measures were 0.646 (QWK), 42.75 (Close) and 22.25 (Correct).

### 5.4 System Details

We conduct our experiments using well-established norms in eye-tracking research (Holmqvist et al., 2011). The essays are displayed on a screen that is kept about 2 feet in front of the participant.

The work flow of the experiment is as follows. First, the camera is calibrated. This is done by having the annotator look at 13 points on the screen, while the camera tracks their eyes. Next, the calibration is validated. In this step, the participant looks at the same points they saw earlier. If there is a big difference between the participant's fixation points tracked by the camera and the actual points, calibration is repeated. Then, the reader

[5] Another advantage of using source-dependent essays is that there is a source article which we can use to correctly replace the anonymized named entities

[6] We report details on individual annotators in Appendix B.

performs a self-paced reading of the essay, while we supervise the tracking of their eyes. After reading and scoring an essay, the participant takes a small break of about a minute, before continuing. Before the next essay is read, the camera has to again be calibrated and validated.[7] The essay is displayed on the screen in Times New Roman typeface with a font size of 23. Finally, the reader scores the essay and provides a justification for their score[8].

This entire process is done using an SR Research Eye Link 1000 eye-tracker (monocular stabilized head mode, with a sampling rate of 500Hz). The machine collects all the gaze details that we need for our experiments. An interest area report is generated for gaze behaviour using the SR Research Data Viewer software.

### 5.5 Experiment Details

We use five-fold cross-validation to evaluate our system. For each fold, 60% is used as training, 20% for validation, and 20% for testing. The folds are the same as those used by Taghipour and Ng (2016). Prior to running our experiments, we convert the scores from their original score range (given in Table 1) to the range of $[0, 1]$ as described by Taghipour and Ng (2016).

In order to normalize idiosyncratic reading patterns across different readers, we perform binning for each of the features for each of the readers. For IR and Skip we use only two bins - 0 and 1 - corresponding to their values. For the run count, we use six bins (from 0 to 5), where each bin is the run count (up to 4), and bin 5 contains run counts more than 4. For the fixation attributes - DT and FFD - we use the same binning scheme as described in Klerke et al. (2016). The binning scheme for fixation attributes is as follows:

$$
\begin{cases}
0 & \text{if } FV = 0, \\
1 & \text{if } FV > 0 \text{ and } FV \leq \mu, \\
2 & \text{if } FV > \mu \text{ and } FV \leq \mu + 0.5\sigma, \\
3 & \text{if } FV > \mu + 0.5\sigma \text{ and } FV \leq \mu + \sigma, \\
4 & \text{if } FV > \mu + \sigma \text{ and } FV \leq \mu + 2\sigma, \\
5 & \text{if } FV > \mu + 2\sigma,
\end{cases}
$$

where $FV$ is the value of the given fixation attribute, $\mu$ is the average fixation attribute value for the reader and $\sigma$ is the standard deviation.

### 5.6 Network Architecture

Figure 1 (b) shows the architecture of our proposed system, based on the co-attention based architecture described by Zhang and Litman (2018). Given an essay, we split the essay into sentences. For each sentence, we look-up the word embeddings for all words in the Word Embedding layer. The 4000 most frequent words are used as the vocabulary, with all other words mapped to a special unknown token. This sequence of word embeddings is then sent through a Time-Delay Neural Network (TDNN), or 1-d Convolutional Neural Network (CNN), of filter width $k$. The output from CNN is pooled using an attention layer - the Word Level Attention Pooling Layer - which results in a representation for every sentence. These sentence representations are then sent through a Sentence Level LSTM Layer and their output pooled in the Sentence Level Attention Pooling Layer to obtain the sentence representation for the essay.

A similar procedure is repeated for the source article. We then perform co-attention between the sentence representations of the essay and the source article. Co-attention is performed to learn similarities between the sentences in the essay and the source article. This is done as a way to ensure that the writer sticks to answering the prompt, rather than drifting off topic.

We now represent every sentence in the essay as a weighted combination of the sentence representation between the essay and the source article (Essay2Article). The weights are obtained from the output of the co-attention layer. The weights represent how each sentence in the essay are similar to the sentences in the source article. If a sentence in the essay has low weights this indicates that the sentence would be off topic. A similar procedure is repeated to get a weighted representation of sentences in the source article with respect to the essay (Article2Essay).

Finally, we send the sentence representation of the essay and article, through a dense layer (i.e. the Modeling Layer) to predict the final essay score, with a sigmoid activation function. As the essay scores are in the range $[0, 1]$, we use sigmoid activation at the output layer. During prediction, we map the output scores from the sigmoid layer back to the original score range, minimizing the mean squared error (MSE) loss.

[7] The average time for the participants was about 2 hours, with the fastest completing the task in slightly under one and a half hours.

[8] As part of our data release, we will release the scores given by each annotator, as well as their justifications for their score.

Figure 1: Architecture of the proposed gaze behaviour and essay scoring multi-task learning systems, namely (a) - the Self-Attention multi-task learning system, for an essay of $n$ sentences - and (b) - the Co-Attention system for an essay of $n$ sentences and a source article of $m$ sentences.

For essay sets without a source article, we use the Self-Attention model proposed by Dong et al. (2017). This is a simpler model which does not consider the source article, and uses only the essay text. This is applicable whenever a source article is not present. Figure 1 (a) shows the architecture of the model. Like the earlier system, we get the sentence representation of the essay from the Sentence Level LSTM Layer and send it through the Dense Layer with a sigmoid activation function.

Gaze behaviour is learnt at the Word-Level Convolutional Layer in both the models because the gaze attributes are defined at the word-level, while the essay is scored at the document-level. The output from the CNN layer is sent through a linear layer followed by sigmoid activation for a particular gaze behaviour. For learning multiple gaze attributes simultaneously, we have multiple linear layers for each of the gaze attributes. In the multi-task setting, we also minimize the mean squared error of the learnt gaze behaviour and the actual gaze behaviour attribute value. We assign weights to each of the gaze behaviour loss functions to control the importance given to individual gaze behaviour learning tasks.

### 5.7 Network Hyperparameters

Table 3 gives the different hyperparameters which we used in our experiment. We use the 50 dimension GloVe pre-trained word embeddings (Pennington et al., 2014) trained on the Wikipedia 2014 + Gigawords 5 Corpus (6B tokens, 4K vocabulary, uncased). We run our experiments over a batch size of 100, for 100 epochs, and set the learning rate as 0.001, and a dropout rate of 0.5. The Word-level CNN layer has a kernel size of 5, with 100 filters. The Sentence-level LSTM layer and modeling layer both have 100 hidden units. We use the RMSProp Optimizer (Dauphin et al., 2015) with a 0.001 initial learning rate and momentum of 0.9.

| Layer | Hyperparameter | Value |
|---|---|---|
| Embedding layer | Pre-trained embeddings | GloVe |
| | Embeddings dimensions | 50 |
| Word-level CNN | Kernel size | 5 |
| | Filters | 100 |
| Sentence-level LSTM | Hidden units | 100 |
| Network-wide | Batch size | 100 |
| | Epochs | 100 |
| | Learning rate | 0.001 |
| | Dropout rate | 0.5 |
| | Momentum | 0.9 |

Table 3: Hyperparameters for our experiment.

| Gaze Feature | Gaze Feature Weight |
|---|---|
| Dwell Time | 0.05 |
| First Fixation Duration | 0.05 |
| IsRegression | 0.01 |
| Run Count | 0.01 |
| Skip | 0.1 |

Table 4: This table shows the best weights assigned to the different gaze features from our grid search.

In addition to the network hyper-parameters, we also weigh the loss functions of the different gaze

behaviours differently, with weight levels of 0.5, 0.1, 0.05, 0.01 and 0.001. We use grid search and pick the weight giving the lowest mean-squared error on the development set. The best weights from grid search are 0.05 for DT and FFD, 0.01 for IR and RC, and 0.1 for Skip.

### 5.8 Experiment Configurations

To test our system on essay sets which we collected gaze behaviour, we run experiments using the following configurations. (a) Self-Attention - This is the implementation of Dong et al. (2017)'s system in Tensorflow by Zhang and Litman (2018). (b) Co-Attention. This is Zhang and Litman (2018)'s system[9]. (c) Co-Attention+Gaze. This is our system, which uses gaze behaviour.

In addition to this, we also run experiments on the unseen essay sets using the following training configurations. (a) Only Prompt - This uses our self-attention model, with the training data being only the essays from that essay set. We use this model, because there are no source articles for these essay sets. (b) Extra Essays - Here, we augment the training data of (a) with the 48 essays for which we collect gaze behaviour data. (c) Essays+Gaze - Here, we augment the training data of (a) with the 48 essays which we collect gaze behaviour data, and their corresponding gaze data. We also compare our results with a string kernel based system proposed by Cozma et al. (2018).

## 6 Results and Analysis

Table 5 reports the results of our experiments on the essay sets for which we collect the gaze behaviour data. The table is divided into 3 parts. The first part (i.e., first 3 rows) are the reported results previously available deep-learning systems, namely Taghipour and Ng (2016), Dong and Zhang (2016), and Tay et al. (2018). The next 2 rows feature results using the self-attention (Dong et al., 2017) and co-attention (Zhang and Litman, 2018). The last row reports results using gaze behaviour on top of co-attention, i.e., Co-Attention+Gaze. The first column is the different systems. The next 4 columns report the QWK results of each system for each of the 4 essay sets. The last column reports the Mean QWK value across all 4 essay sets.

Our system is able to outperform the Co-Attention system (Zhang and Litman, 2018) in all the essay sets. Overall, it is also the best system - achieving the highest QWK results among all the systems in 3 out of the 4 essay sets (and the second-best in the other essay set). To test our hypothesis - that the model trained by learning gaze behaviour helps in automatic essay grading - we run the Paired T-Test. Our null hypothesis is: "Learning gaze behaviour to score an essay does not help any more than the self-attention and co-attention systems and whatever improvements we see are due to chance." We choose a significance level of $p < 0.05$, and observe that the improvements of our system are found to be statistically significant - rejecting the null hypothesis.

### 6.1 Results for Unseen Essay Sets

In order to run our experiments on unseen essay sets, we augment the training data with the gaze behaviour data collected. Since none of these essays have source articles, we use the self-attention model of Dong et al. (2017) as the baseline system. We now augment the gaze behaviour learning task as the auxiliary task and report the results in Table 6. The first column in the table is the different systems. The next 4 columns are the results for each of the unseen essay sets, and the last column is the mean QWK. From Table 6, we observe that our system which uses both the extra 48 essays and their gaze behaviour outperforms the other 2 configurations (Only Prompt and Extra Essays) across all 4 unseen essay sets. The improvement when learning gaze behaviour for unseen essay sets is statistically significant for $p < 0.05$.

### 6.2 Comparison with String Kernel System

Since Cozma et al. (2018) haven't released their data splits (train/test/dev), we ran their system with our data splits. We observed a mean QWK of 0.750 with the string kernel-based system on the essay sets where we have gaze behaviour data, and 0.665 on the unseen essay sets. One possible reason for this could be that while they used cross-validation, they may have used only a training-testing split (as compared to a train/test/dev split).

### 6.3 Analysis of Gaze Attributes

In order to see which of the gaze attributes are the most important, we ran ablation tests, where we ablate each gaze attribute. We found that the most important gaze behaviour attribute across all the essay sets is the Dwell Time, followed closely by the First Fixation Duration. One of the reasons

[9] The implementation of both systems can be downloaded from here.

| System | Prompt 3 | Prompt 4 | Prompt 5 | Prompt 6 | Mean QWK |
|---|---|---|---|---|---|
| Taghipour and Ng (2016) | 0.683 | 0.795 | 0.818 | 0.813 | 0.777 |
| Dong and Zhang (2016) | 0.662 | 0.778 | 0.800 | 0.809 | 0.762 |
| Tay et al. (2018) | 0.695 | 0.788 | 0.815 | 0.810 | 0.777 |
| Self-Attention (Dong et al., 2017) | 0.677 | 0.807 | 0.806 | 0.809 | 0.775 |
| Co-Attention (Zhang and Litman, 2018) | 0.689† | 0.809† | 0.812† | 0.813† | 0.780† |
| Co-Attention+Gaze | 0.698* | 0.818* | 0.815* | 0.821* | 0.788* |

Table 5: Results of our experiments in scoring the essays (QWK values) from the essay sets where we collected gaze behaviour. The first 3 rows are results reported from other state-of-the-art deep learning systems. The next 2 rows are the results we obtained on existing systems - self-attention and co-attention - without gaze behaviour. The last row is the results from our system using gaze behaviour data (Co-Attention+Gaze). † denotes the baseline system performance, and * denotes a statistically significant result with $p < 0.05$ for the gaze behaviour system.

| System | Prompt 1 | Prompt 2 | Prompt 7 | Prompt 8 | Mean QWK |
|---|---|---|---|---|---|
| Taghipour and Ng (2016) | 0.775 | 0.687 | 0.805 | 0.594 | 0.715 |
| Dong and Zhang (2016) | 0.805 | 0.613 | 0.758 | 0.644 | 0.705 |
| Tay et al. (2018) | 0.832 | 0.684 | 0.800 | 0.697 | 0.753 |
| Only Prompt (Dong et al. (2017)) | 0.816 | 0.667 | 0.792 | 0.678 | 0.738 |
| Extra Essays | 0.828† | 0.672† | 0.802† | 0.685† | 0.747† |
| Extra Essays + Gaze | 0.833 | 0.681 | 0.806* | 0.699* | 0.754* |

Table 6: Results of our experiments on the unseen essay sets our dataset. The first 3 rows are results reported from other state-of-the-art deep learning systems. The next 2 rows are the results obtained without using gaze behaviour (without and with the extra essays). The last row is the results from our system. † denotes the baseline system without gaze behaviour, and * denotes a statistically significant result with $p < 0.05$ for the gaze behaviour system.

| Gaze Feature | Diff. in QWK |
|---|---|
| Dwell Time | 0.0137 |
| First Fixation Duration | 0.0136 |
| IsRegression | 0.0090 |
| Run Count | 0.0110 |
| Skip | 0.0091 |

Table 7: Results of ablation tests for each gaze behaviour attribute across all the essay sets. The reported numbers are the difference in QWK before and after ablating the given gaze attribute. The number in bold denotes the best gaze attribute.

for this is the fact that both DT and FFD were very useful in detecting errors made by the essay writers. From Figure 2[10], we observe that most of the longest dwell times have come at/around spelling mistakes (lock instead of took), or out-of-context words (lay instead of by), or incorrect phrases (short cat, instead of short cut). These errors force the reader to spend more time fixating on the word which we also mentioned earlier.

[10] We have given more examples in Appendix C.

The normalized MSE of each of the gaze features learnt by our system was between 0.125 to 0.128 for all the gaze behaviour attributes.

### 6.4 Analysis Using Only a Native English Speaker

| System | No | Native | All |
|---|---|---|---|
| Prompt 1 | 0.816 | 0.824 | 0.833 |
| Prompt 2 | 0.667 | 0.679 | 0.681 |
| Prompt 3 | 0.677 | 0.679 | 0.698 |
| Prompt 4 | 0.807 | 0.812 | 0.818 |
| Prompt 5 | 0.806 | 0.810 | 0.815 |
| Prompt 6 | 0.809 | 0.815 | 0.821 |
| Prompt 7 | 0.792 | 0.809 | 0.806 |
| Prompt 8 | 0.678 | 0.679 | 0.699 |
| Mean QWK | 0.757 | 0.764 | 0.771 |

Table 8: Result using only gaze behaviour of the native speaker (Native), compared using no gaze behaviour (No) and gaze behaviour of all the readers (All).

We also ran our experiments using only the gaze behaviour of an annotator who was a native En-

Figure 2: Dwell Time of one of the readers for one of the essays. The darker the background, the larger the bin.

glish speaker (as opposed to the rest of our annotators who were just uent English speakers). Table 8 shows the results of those experiments. We observed a mean QWK of 0.779 for the seen essay sets, and a mean QWK of 0.748 for the essays sets where we have no gaze data. The difference in performance between both our systems (i.e. with only native speaker and with all annotators) were found to be statistically signi cant with $p < 0.0245$[11]. Similarly, the improvement in performance using the native English speaker, compared to not using any gaze behaviour was also found to be statistically signi cant for $p < 0.0084$.

## 7 Conclusion and Future Work

In this paper, we describe how learning gaze behaviour can help AEG in a multi-task learning setup. We explained how we created a resource by collecting gaze behaviour data, and using multi-task learning we are able to achieve better results over a state-of-the-art system developed by Zhang and Litman (2018) for the essay sets which we collected gaze behaviour data from. We also analyze the transferability of gaze behaviour patterns across essay sets by training a multi-task learning model on unseen essay sets (i.e. essay sets where we have no gaze behaviour data), thereby establishing that learning gaze behaviour improves automatic essay grading.

In the future, we would like to look at using gaze behaviour to help in cross-domain AEG. This is done mainly when we don't have enough training examples in our essay set. We would also like to explore the possibility of generating textual feedback (rather than just a number, denoting the score of the essay) based on the justi cations that the annotators gave for their grades.

[11] The p-values for the different experiments are in Appendix D.

## References


Maria Barrett, Joachim Bingel, Nora Hollenstein, Marek Rei, and Anders Søgaard. 2018. Sequence classi cation with human attention. In Proceedings of the 22nd Conference on Computational Natural Language Learning, pages 302–312, Brussels, Belgium. Association for Computational Linguistics.

Maria Barrett, Joachim Bingel, Frank Keller, and Anders Søgaard. 2016. Weakly supervised part-of-speech tagging using eye-tracking data. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 579–584, Berlin, Germany. Association for Computational Linguistics.

Rich Caruana. 1998. Multitask Learning, pages 95–133. Springer US, Boston, MA.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological bulletin, 70(4):213.

Madalina Cozma, Andrei Butnaru, and Radu Tudor Ionescu. 2018. Automated essay scoring with string kernels and word embeddings. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 503–509, Melbourne, Australia. Association for Computational Linguistics.

Yann Dauphin, Harm De Vries, and Yoshua Bengio. 2015. Equilibrated adaptive learning rates for non-convex optimization. In Advances in neural information processing systems, pages 1504–1512.

Fei Dong and Yue Zhang. 2016. Automatic features for essay scoring – an empirical study. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1072–1077, Austin, Texas. Association for Computational Linguistics.

Fei Dong, Yue Zhang, and Jie Yang. 2017. Attention-based recurrent convolutional neural network for automatic essay scoring. In Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), pages 153–162, Vancouver, Canada. Association for Computational Linguistics.

Ana V González-Garduño and Anders Søgaard. 2018. Learning to predict readability using eye-movement data from natives and learners. In Thirty-Second AAAI Conference on Arti cial Intelligence

Ana Valeria González-Garduño and Anders Søgaard. 2017. Using gaze to predict text readability. In Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, pages 438–443, Copenhagen, Denmark. Association for Computational Linguistics.

John Hale, Chris Dyer, Adhiguna Kuncoro, and Jonathan Brennan. 2018. Finding syntax in human encephalography with beam search. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2727–2736, Melbourne, Australia. Association for Computational Linguistics.

Nora Hollenstein and Ce Zhang. 2019. Entity recognition at first sight: Improving NER with eye movement information. 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 1–10, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Marcel A Just and Patricia A Carpenter. 1980. A theory of reading: From eye fixations to comprehension. Psychological review, 87(4):329.

Alan Kennedy, Robin Hill, and Joël Pynte. 2003. The dundee corpus. In Proceedings of the 12th European conference on eye movement.

Sigrid Klerke, Yoav Goldberg, and Anders Søgaard. 2016. Improving sentence compression by learning to predict gaze. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1528–1533, San Diego, California. Association for Computational Linguistics.

Yunfei Long, Rong Xiang, Qin Lu, Chu-Ren Huang, and Minglei Li. 2019. Improving attention model based on cognition grounded data for sentiment analysis. IEEE Transactions on Affective Computing.

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.

Abhijit Mishra and Pushpak Bhattacharyya. 2018. Cognitively Inspired Natural Language Processing: An Investigation Based on Eye-tracking. Springer.

Abhijit Mishra, Diptesh Kanojia, and Pushpak Bhattacharyya. 2016. Predicting readers' sarcasm understandability by modeling gaze behavior.

Abhijit Mishra, Diptesh Kanojia, Seema Nagar, Kuntal Dey, and Pushpak Bhattacharyya. 2017. Scanpath complexity: Modeling reading effort using gaze information.

Abhijit Mishra, Srikanth Tamilselvam, Riddhiman Dasgupta, Seema Nagar, and Kuntal Dey. 2018. Cognition-cognizant sentiment analysis with multitask subjectivity summarization based on annotators' gaze behavior. In Thirty-Second AAAI Conference on Artificial Intelligence.

Ellis B Page. 1966. The imminence of... grading essays by computer. The Phi Delta Kappan, 47(5):238–243.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Mark D Shermis and Jill Burstein. 2013. Handbook of automated essay evaluation: Current applications and new directions. Routledge.

Abhinav Deep Singh, Poojan Mehta, Samar Husain, and Rajkumar Rajakrishnan. 2016. Quantifying sentence complexity based on eye-tracking measures. In Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC), pages 202–212, Osaka, Japan. The COLING 2016 Organizing Committee.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1882–1891, Austin, Texas. Association for Computational Linguistics.

Yi Tay, Minh Phan, Luu Anh Tuan, and Siu Cheung Hui. 2018. Skip flow: Incorporating neural coherence features for end-to-end automatic text scoring.

Haoran Zhang and Diane Litman. 2018. Co-attention based neural network for source-dependent essay scoring. In Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 399–409, New Orleans, Louisiana. Association for Computational Linguistics.

## A Source Article (Prompt 6)

The Mooring Mast, by Marcia Amidon Lüsted

When the Empire State Building was conceived, it was planned as the world's tallest building, taller even than the new Chrysler Building that was being constructed at Forty-second Street and Lexington

Avenue in New York. At seventy-seven stories, it was the tallest building before the Empire State began construction, and Al Smith was determined to outstrip it in height.

The architect building the Chrysler Building, however, had a trick up his sleeve. He secretly constructed a 185-foot spire inside the building, and then shocked the public and the media by hoisting it up to the top of the Chrysler Building, bringing it to a height of 1,046 feet, 46 feet taller than the originally announced height of the Empire State Building.

Al Smith realized that he was close to losing the title of world's tallest building, and on December 11, 1929, he announced that the Empire State would now reach the height of 1,250 feet. He would add a top or a hat to the building that would be even more distinctive than any other building in the city. John Tauranac describes the plan:

"[The top of the Empire State Building] would be more than ornamental, more than a spire or dome or a pyramid put there to add a desired few feet to the height of the building or to mask something as mundane as a water tank. Their top, they said, would serve a higher calling. The Empire State Building would be equipped for an age of transportation that was then only the dream of aviation pioneers."

This dream of the aviation pioneers was travel by dirigible, or zeppelin, and the Empire State Building was going to have a mooring mast at its top for docking these new airships, which would accommodate passengers on already existing transatlantic routes and new routes that were yet to come.

## A.1 The Age of Dirigibles

By the 1920s, dirigibles were being hailed as the transportation of the future. Also known today as blimps, dirigibles were actually enormous steel-framed balloons, with envelopes of cotton fabric filled with hydrogen and helium to make them lighter than air. Unlike a balloon, a dirigible could be maneuvered by the use of propellers and rudders, and passengers could ride in the gondola, or enclosed compartment, under the balloon.

Dirigibles had a top speed of eighty miles per hour, and they could cruise at seventy miles per hour for thousands of miles without needing refueling. Some were as long as one thousand feet, the same length as four blocks in New York City. The one obstacle to their expanded use in New York City was the lack of a suitable landing area. Al Smith saw an opportunity for his Empire State Building: A mooring mast added to the top of the building would allow dirigibles to anchor there for several hours for refueling or service, and to let passengers off and on. Dirigibles were docked by means of an electric winch, which hauled in a line from the front of the ship and then tied it to a mast. The body of the dirigible could swing in the breeze, and yet passengers could safely get on and off the dirigible by walking down a gangplank to an open observation platform.

The architects and engineers of the Empire State Building consulted with experts, taking tours of the equipment and mooring operations at the U.S. Naval Air Station in Lakehurst, New Jersey. The navy was the leader in the research and development of dirigibles in the United States. The navy even offered its dirigible, the Los Angeles, to be used in testing the mast. The architects also met with the president of a recently formed airship transport company that planned to offer dirigible service across the Paci c Ocean.

When asked about the mooring mast, Al Smith commented:

"[It's] on the level, all right. No kidding. We're working on the thing now. One set of engineers here in New York is trying to dope out a practical, workable arrangement and the Government people in Washington are guring on some safe way of mooring airships to this mast."

## A.2 Designing the Mast

The architects could not simply drop a mooring mast on top of the Empire State Building's at roof. A thousand-foot dirigible moored at the top of the building, held by a single cable tether, would add stress to the building's frame. The stress of the dirigible's load and the wind pressure would have to be transmitted all the way to the building's foundation, which was nearly eleven hundred feet below. The steel frame of the Empire State Building would have to be modi ed and strengthened to accommodate this new situation. Over sixty thousand dollars' worth of modi cations had to be made to the building's framework.

Rather than building a utilitarian mast without any ornamentation, the architects designed a shiny glass and chrome-nickel stainless steel tower that would be illuminated from inside, with a stepped-back design that imitated the overall shape of the

building itself. The rocket-shaped mast would have four wings at its corners, of shiny aluminum, and would rise to a conical roof that would house the mooring arm. The winches and control machinery for the dirigible mooring would be housed in the base of the shaft itself, which also housed elevators and stairs to bring passengers down to the eighty-sixth floor, where baggage and ticket areas would be located.

The building would now be 102 floors, with a glassed-in observation area on the 101st floor and an open observation platform on the 102nd floor. This observation area was to double as the boarding area for dirigible passengers.

Once the architects had designed the mooring mast and made changes to the existing plans for the building's skeleton, construction proceeded as planned. When the building had been framed to the 85th floor, the roof had to be completed before the framing for the mooring mast could take place. The mast also had a skeleton of steel and was clad in stainless steel with glass windows. Two months after the workers celebrated framing the entire building, they were back to raise an American flag again—this time at the top of the frame for the mooring mast.

### A.3 The Fate of the Mast

The mooring mast of the Empire State Building was destined to never fulfill its purpose, for reasons that should have been apparent before it was ever constructed. The greatest reason was one of safety: Most dirigibles from outside of the United States used hydrogen rather than helium, and hydrogen is highly flammable. When the German dirigible Hindenburg was destroyed by fire in Lakehurst, New Jersey, on May 6, 1937, the owners of the Empire State Building realized how much worse that accident could have been if it had taken place above a densely populated area such as downtown New York.

The greatest obstacle to the successful use of the mooring mast was nature itself. The winds on top of the building were constantly shifting due to violent air currents. Even if the dirigible were tethered to the mooring mast, the back of the ship would swivel around and around the mooring mast. Dirigibles moored in open landing fields could be weighted down in the back with lead weights, but using these at the Empire State Building, where they would be dangling high above pedestrians on the street, was neither practical nor safe.

The other practical reason why dirigibles could not moor at the Empire State Building was an existing law against airships flying too low over urban areas. This law would make it illegal for a ship to ever tie up to the building or even approach the area, although two dirigibles did attempt to reach the building before the entire idea was dropped. In December 1930, the U.S. Navy dirigible Los Angeles approached the mooring mast but could not get close enough to tie up because of forceful winds. Fearing that the wind would blow the dirigible onto the sharp spires of other buildings in the area, which would puncture the dirigible's shell, the captain could not even take his hands off the control levers.

Two weeks later, another dirigible, the Goodyear blimp Columbia, attempted a publicity stunt where it would tie up and deliver a bundle of newspapers to the Empire State Building. Because the complete dirigible mooring equipment had never been installed, a worker atop the mooring mast would have to catch the bundle of papers on a rope dangling from the blimp. The papers were delivered in this fashion, but after this stunt the idea of using the mooring mast was shelved. In February 1931, Irving Clavan of the building's architectural office said, “The as yet unsolved problems of mooring air ships to a fixed mast at such a height made it desirable to postpone to a later date the final installation of the landing gear.”

By the late 1930s, the idea of using the mooring mast for dirigibles and their passengers had quietly disappeared. Dirigibles, instead of becoming the transportation of the future, had given way to airplanes. The rooms in the Empire State Building that had been set aside for the ticketing and baggage of dirigible passengers were made over into the world's highest soda fountain and tea garden for use by the sightseers who flocked to the observation decks. The highest open observation deck, intended for disembarking passengers, has never been open to the public.

## B Annotator Profiles

Table 9 summarizes the profiles of the different annotators. It details each of the 8 annotators, their sex, age, occupations, L1 / native languages, their performance in a high school Examination in English and whether or not they have had experience as a TA. The last 3 columns are their performance

| ID | Sex | Age | Occupation | TA? | L1 Language | English Score | QWK | Correct | Close |
|---|---|---|---|---|---|---|---|---|---|
| Annotator 1 | Male | 23 | Masters student | Yes | Hindi | 94% | 0.611 | 19 | 41 |
| Annotator 2 | Male | 18 | Undergraduate | Yes | Marathi | 95% | 0.587 | 24 | 41 |
| Annotator 3 | Male | 31 | Research scholar | Yes | Marathi | 85% | 0.659 | 21 | 43 |
| Annotator 4 | Male | 28 | Software engineer | Yes | English | 96% | 0.659 | 26 | 44 |
| Annotator 5 | Male | 30 | Research scholar | Yes | Gujarati | 92% | 0.600 | 19 | 42 |
| Annotator 6 | Female | 22 | Masters student | Yes | Marathi | 95% | 0.548 | 19 | 40 |
| Annotator 7 | Male | 19 | Undergraduate | Yes | Marathi | 93% | 0.732 | 21 | 46 |
| Annotator 8 | Male | 28 | Masters student | Yes | Gujarati | 94% | 0.768 | 29 | 45 |

Table 9: Pro le of the annotators

on the annotation grading task, where QWK is their agreement with the ground truth scores, Correct is the number of times (out of 48) where their essay scores matched with the ground truth scores, and Close is the number of times (out of 48) where they disagreed with the ground truth score by at most 1 grade point.

# C Heat Map Examples

## C.1 Different Gaze Features

Here, we show examples of heat maps for different gaze behaviour attributes of one of our readers.

1. Figure 3 shows the dwell time of the reader.
2. Figure 4 shows the heat map of the rst xation duration of a reader.
3. Figure 5 shows the heat map of the IsRegression feature - i.e. whether or not the reader regressed from a particular word.
4. Figure 6 shows the heat map of the Run Count of the reader.
5. Figure 7 shows the words that the reader read (highlighted) and skipped (unhighlighted).

## C.2 Dwell Times of Good and Bad Essays

Figures 8 and 9 show the dwell time heat maps of a reader as he reads a good essay and a bad essay respectively. For the bad essay, notice the amount of a lot more darker blues compared to the good essay.

# D P-Values

In this section, we report the p-values and other results for our experiments.

## D.1 Source-Dependent Essay Set's p-values

The results shown here in Table 10 are the p-values for the different essay sets with and without gaze from Table 5.

| Essay Set | p-value |
|---|---|
| Prompt 3 | 0.0042 |
| Prompt 4 | 0.0109 |
| Prompt 5 | 0.0133 |
| Prompt 6 | 0.0003 |

Table 10: Source-Dependent essay set's p-values

## D.2 Unseen Essay Set's p-values

The results shown here in Table 10 are the p-values for the different essay sets with and without gaze from Table 6.

| Essay Set | p-value |
|---|---|
| Prompt 1 | 0.0887 |
| Prompt 2 | 0.1380 |
| Prompt 7 | 0.0393 |
| Prompt 8 | 0.0315 |

Table 11: Unseen Essay's p-values

## D.3 Native Gaze vs. No Gaze & All Gaze p-values

The results shown in Table 12 are the p-values for the essay sets using the gaze behaviour of a native English speaker compared to not using gaze behaviour, and using gaze behaviour of all readers.

| Essay Set | No vs. Native | Native vs. All |
|---|---|---|
| Prompt 1 | 0.1407 | 0.0471 |
| Prompt 2 | 0.0161 | 0.9161 |
| Prompt 3 | 0.3239 | 0.0239 |
| Prompt 4 | 0.0810 | 0.0805 |
| Prompt 5 | 0.4971 | 0.4010 |
| Prompt 6 | 0.2462 | 0.2961 |
| Prompt 7 | 0.0189 | 0.0098 |
| Prompt 8 | 0.8768 | 0.0068 |

Table 12: No gaze vs. native gaze and native gaze vs. all gaze p-values.

Figure 3: Sample heat map of the dwell of a reader for the text. The darker the blue, the larger the bin, and the longer the dwell time.

Figure 4: Sample heat map of the rst xation duration of a reader for the text. The darker the blue, the larger the bin, and the longer the rst xation duration.

Figure 5: Sample heat map of the Is Regression feature of a reader for the text. The highlighted words denote words that the reader regressed from.

Figure 6: Sample heat map of the run count of a reader for the text. The darker the blue, the larger the bin, and the higher the run count.

Figure 7: Sample heat map of the Skip feature of a reader for the texturThe highlighted words denote words that the reader skipped.