Journal of
Eye Movement Research

MDPI

Article

Reader Performance History as a Benchmark for Personalised Reading Comprehension Prediction from Eye Movements: Evidence from OneStop, SB-SAT, and PoTeC

Akshay Mendhakar * and Kirren Chana

Department of Cognition, Emotion, and Methods in Psychology, University of Vienna, 1010 Vienna, Austria
* Correspondence: akshay.mendhakar@univie.ac.at

Abstract

Models of reading comprehension use eye movements to predict trial-level comprehension. However, previous labelled responses may also provide information about a reader's expected performance. We tested this using OneStop Eye Movements as the primary dataset, with exploratory external evaluations in SB-SAT and PoTeC. The planned temporal-gaze analysis did not support the hypothesised monotonic increase in predictive value as more fixations became available. Subsequent exploratory analyses evaluated reader performance history as a baseline for personalised prediction. In a matched training-only EyeBench OneStop ordinary-reading, seen-reader/unseen-text comparison (7774 trials), empirical-Bayes reader history achieved AUROC of 0.6101 versus 0.5785 for MAG-Eye (paired t-test $p = 0.0444$; exact sign-flip $p = 0.0508$). Reader-history performance increased with more labelled responses, and reader and item histories provided complementary predictive information. In information seeking, matched whole-article holdout analyses showed that positive contribution of gaze over text was reduced after including reader history, both at 20 fixations and for the full record (exact sign-flip $p = 0.016$ and 0.006 respectively). Exploratory external comparisons were inconclusive because of wide uncertainty. These findings support evaluating reader performance history as a baseline for personalised comprehension prediction and suggest that part of the predictive contribution of gaze reflects stable differences between readers.

Keywords: eye tracking; reading comprehension; reader history; individual differences

Check for updates

Academic Editors: Tanya R. Beelders and Anne E. Cook

Received: 12 October 2026
Revised: 18 November 2026
Accepted: 20 November 2026
Published: 24 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Reading comprehension is widely understood as a complex process through which readers construct meaning in interaction with the text, their prior knowledge, the purposes of reading and the broader context in which reading occurs [1]. Contemporary accounts therefore treat comprehension as arising not only from properties of the text or characteristics of the reader but from the dynamic relationship between these factors. Construction-integration approaches emphasise that readers build and update a mental representation by integrating textual information with prior knowledge [2]. The Reading Systems Framework places word knowledge and language comprehension within a broader architecture of reader and text factors [3]. Similarly, the RAND Reading Study Group framework defines comprehension as the simultaneous interaction of reader, text, and activity [1].

This interactional view is especially important for prediction systems that estimate whether a reader will answer a comprehension question correctly. At least three sources

of variation may contribute to such outcomes, including reading ability, the difficulty of the text or question and the variation in how the reader processes the material during the current reading task. A predictive system that uses behavioural measures should therefore establish which of these sources of information it captures. This distinction becomes especially important when eye movements are used to predict comprehension.

1.1. Eye Movements and Reading Comprehension

During reading, eye-tracking provides a temporally resolved measure of overt visual attention [4]. The eye–mind assumption suggests that fixation behaviour relates to ongoing cognitive processing [5]. Decades of research have established systematic sensitivity of eye movements to lexical, syntactic, semantic, and discourse-level variables [4,6,7]. Modern computational accounts, such as E-Z Reader and SWIFT, formalise how lexical processing, attention, and oculomotor control jointly generate scanpaths [8,9]. Empirical research has also shown that fixation behaviour can vary across reading platforms [10]. At the same time, eye movements are highly idiosyncratic and can support reader identification [11]. Individual differences in reading skill can affect fixation times and modulate canonical word-length and frequency effects, in some cases more strongly than the lexical predictors themselves [12]. This dual character of gaze as both a process measure and a marker of stable reader differences creates a central identification problem for predictive modelling.

Previous research supports an association between eye movements and subsequent comprehension. In studies with longer texts, global eye-movement measures can predict later comprehension scores across participants. Southwell et al. [13] reported cross-validated correlations of approximately 0.36–0.38 across three datasets and found that more frequent but shorter fixations were consistently associated with better comprehension. Mézière et al. [14] found that eye-tracking measures provided information beyond reading speed across several comprehension tasks, but the most informative gaze measures varied by task. These results support the broad notion that eye movements carry comprehension-related information, but they do not establish that the information is purely trial-specific or that it is independent of stable reader ability.

This distinction matters because eye movements are also highly individual. Scanpaths can contain sufficient person-specific information to support reader identification [11], and individual differences in reading skill can influence fixation durations and modulate established effects of variables such as word length and word frequency [12]. A reader who performs well across comprehension tasks may therefore also show systematic eye-movement patterns across texts. Gaze can consequently provide predictive information for at least two different reasons: because it reflects processing on the current trial, or because it carries information about the reader who produced it. These two sources of information are not necessarily independent.

This creates an identification problem for personalised comprehension prediction. A model may predict comprehension successfully from gaze without the predictive signal being entirely trial-specific. If stable reader characteristics influence both comprehension performance and eye movements, part of the apparent gaze effect may reflect information about who the reader is rather than what is happening during the current reading episode.

1.2. From Gaze Prediction to Reader-Specific Generalisation

Machine-learning studies have made this problem increasingly visible by testing whether gaze-based comprehension models generalise beyond the observations on which they were trained. This reflects the broader distinction between explaining existing data and making predictions that generalise to new data [15].

Ahn et al. [16] used fixation-location, duration, and pupil-size sequences from SB-SAT and reported a best overall comprehension classification accuracy of 65%, compared with a 54% null-accuracy benchmark. Performance generalised to new passages read by the same participants (64%) but fell to 41% for new readers, below the reported 54% null-accuracy benchmark. This contrast suggests that reader-specific characteristics may contribute to gaze-based predictions under the evaluated protocol. It does not, by itself, distinguish stable reader information from trial-specific comprehension processes.

This does not imply that gaze merely encodes reader identity or that online processing is unimportant. Rather, it shows that stable reader information and current-trial processing may be statistically entangled. A model evaluated on familiar readers can exploit both sources, whereas a model evaluated on new readers must rely on patterns that generalise across individuals.

More recent work has moved towards finer-grained, trial-level comprehension prediction (see Supplementary Materials). Shubi et al. [17] examined whether a reader would answer a comprehension question correctly after reading a single paragraph. Using the OneStop corpus, they compared traditional and multimodal models under ordinary and information-seeking reading and under generalisation to new readers, new items, or both. Their results show that the prediction task is difficult and that gains from gaze should be judged against strong text-only baselines rather than against chance alone.

EyeBench [18] now provides a unified benchmark for these questions. It harmonises multiple eye-tracking corpora, implements 15 models, and evaluates three realistic generalisation regimes of unseen reader, unseen text, and unseen reader plus unseen text. For trial-level reading comprehension, EyeBench includes OneStop, SB-SAT, and PoTeC. The distinction between these regimes is important because they represent different prediction problems. A model that predicts comprehension for a familiar reader encountering a new text is solving a personalised prediction problem. A model that must predict comprehension for a completely new reader is instead facing a cold-start problem.

The seen-reader/unseen-text regime is therefore particularly important for understanding personalised gaze-based prediction. In this setting, the reader is already known to the system, even though the test text is new. This means that information from the reader's previous behaviour may, in principle, already be available. Yet this possibility has received comparatively little attention in gaze-based comprehension benchmarks.

1.3. The Missing Baseline in Personalised Prediction

Research on individual differences has established that reading behaviour and comprehension vary with vocabulary, decoding skill, rapid naming, working memory, language experience and prior knowledge [3,12]. Recent work on reader-specific representations likewise suggests that models trained on population-level averages may not describe individual readers' performance equally well [19]. Such differences are particularly relevant for binary comprehension outcomes, where some readers may have a consistently higher probability of answering correctly than others.

In a personalised setting, previous labelled comprehension responses provide a simple way to represent this information. If a reader has answered questions on other texts, these outcomes provide an estimate of that reader's general tendency to answer comprehension questions correctly. We refer to this information as reader performance history. Reader performance history does not assume that previous accuracy measures a single underlying cognitive ability. It may reflect several relatively stable or recurring influences, including language knowledge, reading skill, domain knowledge, strategy, motivation, and test-taking behaviour. Its value here is methodological: it provides a direct baseline for the amount of predictable between-reader variation already available to a personalised system.

This baseline is important because personalised gaze models are often compared with text-based or population-level predictors rather than with the same reader's previous comprehension performance. As a result, successful gaze prediction for known readers is difficult to interpret. A model may perform well because current eye movements capture trial-specific difficulty, but it may also recover stable reader information that could have been estimated more directly from previous responses.

The relevant benchmark question is therefore more demanding than whether gaze simply predicts comprehension. Instead, it is whether current-trial gaze predicts comprehension beyond what is already known about the reader. If gaze provides a substantial improvement after stable reader information has been represented, this strengthens the interpretation that eye movements contain trial-specific information about ongoing comprehension. If the improvement becomes small once reader history is included, some of the apparent value of gaze in personalised settings may instead reflect stable differences between readers.

A second source of stable variation must also be considered. Some comprehension questions or texts are simply more difficult than others. A reader may therefore appear to perform well because they encountered easier items, just as an item may appear easy because it was encountered by stronger readers. Reader history should consequently be interpreted together with information about item difficulty rather than in isolation.

1.4. A Psychometric View of Personalised Comprehension Prediction

Psychometric models provide a useful framework for separating these sources of variation. In the Rasch model, the probability of a correct response is determined by a person ability parameter and an item difficulty parameter on a shared latent scale [20]. The present problem does not require a strict Rasch model, but the person–item distinction provides a useful conceptual starting point.

Applied to personalised reading-comprehension prediction, information from a reader's previous responses can estimate a relatively stable reader propensity, whereas item history or text information can represent differences in question or stimulus difficulty. Current-trial gaze can then contribute a third source of information about how the reader is processing the text at that particular moment. Under this decomposition, the value of gaze becomes more specific: gaze is most informative if it explains variance that remains after reader history and text-based item information have already been represented. The conceptual decomposition is illustrated in Supplementary Figure S1.

This distinction changes the interpretation of multimodal prediction. An improvement over a text-only model does not by itself show that gaze captures moment-to-moment comprehension difficulty. If the text model contains no information about the reader, gaze may partly improve prediction by providing indirect information about stable reader characteristics. A stronger test is therefore to ask whether gaze improves prediction after reader-level information has already been included.

The same reasoning applies to item effects. Reader and item information may each explain substantial but different components of comprehension performance. The remaining predictive problem concerns the interaction between the reader and the particular item, together with temporary states such as attention, difficulty, or changes in reading strategy. These are the components for which current-trial gaze should, in principle, be most informative.

This leads to the central conceptual question of the present study: does current-trial gaze improve comprehension prediction after reader history and text-based item information have already been represented?

1.5. Present Study

The present study examined how much current-trial eye movements contribute to personalised reading-comprehension prediction and whether their predictive contribution remains once reader history and available text-based item information are taken into account. We used OneStop Eye Movements as the primary dataset and SB-SAT and PoTeC as external datasets within the EyeBench framework. OneStop provided a particularly useful primary test case because it contains 19,438 eligible first-reading trials from 360 readers, with trial-level comprehension responses and fixation-level eye-movement data from the same readers across multiple texts. SB-SAT and PoTeC provided exploratory cross-dataset checks in samples that differed in language, text material and experimental structure. The datasets and their respective roles are summarised in Supplementary Table S1.

The study comprised a planned analysis followed by exploratory and validation analyses. The planned temporal-gaze analysis examined whether current-trial gaze improved comprehension prediction beyond text information and whether its incremental predictive contribution increased as progressively more of the fixation sequence became available. We addressed the following research questions:

RQ1: Does current-trial gaze improve comprehension prediction beyond text information and does its incremental predictive contribution increase as progressively more of the fixation sequence becomes available?

We hypothesised that adding current-trial gaze to text information would improve comprehension prediction and that its incremental predictive contribution would increase as longer portions of the fixation sequence were observed. Descriptive patterns across the evaluated reader-generalisation regimes subsequently motivated an exploratory analysis of labelled comprehension outcomes from the same reader on other texts. This analysis examined reader performance history as a representation of stable reader-specific information and addressed two further research questions.

RQ2: In the personalised seen-reader/unseen-text setting, how strongly does a reader's previous comprehension performance predict responses to new texts, and how does this prediction change as additional labelled responses become available?

RQ3: In ordinary reading, does current-trial gaze provide additional predictive information once reader history and text-based item information are represented?

Because reader performance history emerged from exploratory analysis rather than from a planned analysis, RQ2 and RQ3 are treated as exploratory questions rather than as confirmatory directional hypotheses. We subsequently evaluated the reader-history findings through training-only estimation, empirical-Bayes shrinkage, item-difficulty adjustment, few-shot support, prospective chronological prediction, whole-article holdout, repeated-reading validation and external evaluation in SB-SAT and PoTeC. Throughout the paper, we therefore distinguish the planned temporal-gaze analysis from exploratory discovery, subsequent robustness analyses, repeated-reading validation and exploratory cross-dataset evaluation. The design and evidential status of each analysis are summarised in Supplementary Table S2.

2. Materials and Methods

This is a secondary analysis of publicly available data and released benchmark outputs. The analyses drew on data from 360 readers in OneStop, 95 readers in SB-SAT and 75 readers in PoTeC. The corresponding trial counts, comprehension formulations and roles of the three datasets are summarised in Table S1 (see Supplementary Materials). No new participant data were collected for the present study. The analysis sequence is reported explicitly because the main reader-history contribution emerged during exploration. The temporal-gaze analysis was planned before the full-data run. Reader performance history,

empirical-Bayes shrinkage, psychometric decomposition, repeated-reading validation, and cross-dataset evaluation were developed after descriptive gaze patterns across the evaluation regimes motivated a closer examination of stable reader effects (see Section 3.1). We therefore distinguish throughout between planned analyses, exploratory discovery, robustness analyses, and exploratory external evaluation rather than treating all analyses as having the same confirmatory status (see Supplementary Table S2). OpenAI's ChatGPT (GPT-5.6 Sol, August 2026) assisted with analysis-code development, exploratory analysis planning and language editing. The authors reviewed and executed the analysis scripts and verified the reported numerical results against archived trial-level predictions, fold-level results and independent numerical recalculations.

2.1. Datasets and Outcome Measures

We analysed OneStop Eye Movements as the primary dataset and used SB-SAT and PoTeC for exploratory external evaluation within EyeBench. OneStop is uniquely suited to the primary analyses because it combines a large sample, trial-level comprehension questions, ordinary and information-seeking reading regimes, repeated reading, fixation-level records, word-level interest-area measures, and participant metadata [21,22]. EyeBench provides a harmonised reading-comprehension formulation for OneStop, SB-SAT, and PoTeC, and released fold assignments and benchmark predictions [18].

OneStop contains eye movements from 360 native English readers reading newswire articles divided into paragraphs and answering multiple-choice comprehension questions [21,22]. The reading-goal manipulation used a between-participants design: 180 readers completed ordinary reading for comprehension with the question displayed after the text was presented, whereas 180 completed information-seeking reading in which the question was previewed before the paragraph. The corpus also includes repeated-reading trials for a subset of previously encountered articles. The present primary first-reading analyses excluded practice and repeated-reading trials, yielding 19,438 trials: 9718 ordinary-reading trials and 9720 information-seeking trials. Across these trials, 16,340 responses were correct and 3098 were incorrect, corresponding to an overall accuracy of 84.06%.

OneStop used a tower-mounted SR Research EyeLink 1000 Plus (SR Research Ltd., Ottawa, ON, Canada) sampling at 1000 Hz. The dominant eye was recorded, with occasional switching to the non-dominant eye when tracking required it. Participants were instructed to read silently and minimise head movement. Stimuli were presented on a 27-inch Dell U2715H monitor (2560 × 1440 pixels; 60 Hz). Calibration used a nine-point bullseye procedure and was performed at least three times per session, with additional recalibration triggered by drift-sensitive fixation targets. Experimenters repeated calibration until the mean validation error was below 0.3 degrees where possible; the reported mean validation error immediately preceding text reading was 0.24 degrees (SD = 0.05). We analysed the authors' released SR Data Viewer fixation and interest-area reports (Data Viewer 4.3.210) rather than re-segmenting raw samples. Word interest areas corresponded to whitespace-tokenised words, and the released reports included trial metadata and linguistic annotations. This preserves the canonical event segmentation used by the source dataset while making pre-processing provenance explicit [23]. Reporting follows the minimum eye-tracking reporting guideline [24]. These details are reported explicitly because eye-tracking results depend on acquisition, calibration, validation, and pre-processing choices; the reporting structure follows recent recommendations for transparent eye-tracking methods [24,25].

SB-SAT contains eye-tracking data from 95 participants who read four passages drawn from SAT reading-comprehension materials [16]. In EyeBench, page-level recordings are concatenated to form passage-level scanpaths, and each passage–question pair is treated as a trial-level comprehension instance. The original study also provides participant-level

subjective difficulty and aggregate comprehension measures. We used the EyeBench trial labels, fold assignments, and released prediction format so that reader performance history was evaluated under the same unseen-text structure as the benchmark models.

PoTeC is a German naturalistic reading corpus with 75 native German readers who read 12 scientific texts in physics and biology [26]. Participants vary in academic background and domain expertise, and each text is followed by text-comprehension and background-knowledge questions. EyeBench defines 2700 trial-level reading-comprehension instances by pairing each text with its comprehension questions. PoTeC is especially useful as an exploratory external evaluation because it differs from OneStop in language, text genre, participant sample, and experimental structure.

2.2. Primary Outcome and Evaluation Regimes

The target variable was binary correctness for a single reader answering a single comprehension question about a single text or passage. Discrimination was quantified primarily by area under the receiver-operating-characteristic curve (AUROC), with balanced accuracy used as a secondary threshold metric. Probability quality was assessed using log loss, Brier score, calibration intercept and slope, and expected calibration error with ten bins (ECE10). AUROC was selected as the primary cross-dataset metric because class proportions differ by dataset and condition and because EyeBench reports AUROC for classification tasks.

We retained EyeBench's generalisation terminology. The central reader-history question concerns the seen-reader/unseen-text regime: test texts are held out from model fitting, but the reader is represented by labelled outcomes on other texts. This is the natural personalised setting for a reader-history prior. It differs from the unseen-reader regime, in which no labelled history exists without an onboarding phase, and from the unseen-reader + unseen-text regime, which is a population-level cold-start problem.

The reported OneStop reader-history AUROCs correspond to different protocols. The benchmark fold mean of 0.619 uses the empirical-Bayes estimator with training-plus-validation labelled support on the released 8746-trial ordinary-reading evaluation set. The ordinary-reading AUROC of approximately 0.612 refers to separate reader-history and conditional-gaze analyses over 9718 first-reading trials. The whole-article holdout analysis yields a mean reader-history AUROC of 0.605 in ordinary reading and 0.602 in information seeking. These estimates use different splits or support rules and should not be interpreted as alternative estimates of one identical comparison.

The primary OneStop dataset comprised 19,438 eligible first-reading trials: 9718 ordinary-reading trials from 180 readers and 9720 information-seeking trials from a different 180 readers. EyeBench's OneStop reading-comprehension benchmark uses ordinary reading only [18]. We inferred the ten held-out text folds and their seen-reader groups using the archived MAG-Eye test predictions, and then classified every ordinary-reading trial by item fold and reader membership. Each item belonged to one fold, and each fold contained 162 seen readers and 18 withheld readers. This deterministic evaluation identified exactly 8746 seen-reader/unseen-text trials, matching all 8746 paired predictions by trial identifier. The remaining 972 trials were all in the unseen-reader/unseen-text intersection.

2.3. Reader-History Estimators

Reader performance history means labelled comprehension outcomes from the same reader on other passages. The corresponding reader-history prior is a statistical estimate of that reader's baseline probability of answering correctly. We reserve the word calibration for eye-tracker calibration/validation and, separately, for statistical probability calibration. Reader history therefore does not refer to calibrating the eye-tracking device. The primary

reader-history estimator used empirical-Bayes shrinkage. Within each training/support fold, a beta prior with parameters α and β was estimated from the distribution of reader-level performance. For a reader with c correct responses among n labelled support responses, the reader-history estimate was the posterior mean, $(c + \alpha)/(n + \alpha + \beta)$. This shrinks estimates based on sparse histories towards the population mean. Only permitted training/support outcomes were used; a test outcome never contributed to its own reader-history estimate. For the headline EyeBench comparison, reader history used labelled training and validation support. We also report training-only support as a stricter sensitivity analysis.

We also fitted a stricter training-only reader-effect model in which sparse text features and one-hot reader indicators were entered into regularised logistic regression. This analysis does not use validation labels as support. It therefore tests whether a reader effect learned strictly from the model-training split improves prediction relative to the same text representation alone. The reader-only and text + reader variants provide a conservative internal validation that is independent of the released EyeBench model probabilities. The training-only reader-effect results are reported in Supplementary Table S3.

2.4. Few-Shot and Prospective Reader-History Analyses

Few-shot analyses restricted the support set to k labelled responses from each reader. For OneStop, k = 1, 3, 5, 10, 20, and 40 were evaluated, with 30 independent support samples per fold to avoid conclusions based on a convenient draw. PoTeC and SB-SAT used the largest support sizes permitted by their smaller number of trials per reader. Reader performance was compared with the strongest released unseen-text EyeBench comparator in the corresponding dataset. We interpret overlap with the benchmark's performance range descriptively and do not claim formal statistical equivalence unless the paired comparison supports it. The few-shot estimates are reported in Supplementary Table S4 and probability-calibration results for the reader-history estimators are supplied in Supplementary Table S5.

A separate prospective OneStop analysis derived trial order and allowed only genuinely earlier labelled responses to contribute to the reader history for a later test trial. Trial identifiers were canonicalised because the raw OneStop representation encodes Boolean fields as True/False whereas EyeBench encodes them as 1/0. We independently derived trial order from the original OneStop fixation archives using participant identifiers, TRIAL_INDEX and trial-start timestamps, matched the resulting keys to all 19,438 first-reading trials and verified label concordance. After harmonisation, the chronological analysis yielded valid support coverage for approximately 90% of test cases (reported in Supplementary Table S6). This analysis provides the most direct operational interpretation of reader performance history as information that could accumulate during a real session.

2.5. Psychometric Reader-Item Decomposition

To test whether reader performance history merely reflects readers receiving systematically easier support items, we fitted regularised reader + item logistic models analogous to a Rasch decomposition. For reader p and item i, the linear predictor contains a reader parameter and an item parameter. We then evaluated the reader component on unseen items. Regularisation sensitivity was examined over C = 0.01, 0.1, 1, 10, and 100. The stability of reader AUROC across this range was used to assess whether the effect depended on an arbitrary penalty strength. We additionally computed leave-one-out empirical-Bayes reader and population item histories and their combination to compare the predictive contributions of reader and item histories without interpreting AUROC as a decomposition of explained variance.

2.6. Eye-Movement Features and Temporal Prefixes

The original planned analysis represented gaze using leakage-safe features computed from only the information observed by a given horizon. Fixation-count horizons were 5, 10, 20, and 40 fixations, plus the full record. Feature families included cumulative and distributional fixation duration, elapsed observed time, pupil summaries, spatial dispersion, inter-fixation/saccade distance, progression and regression behaviour, unique interest areas reached, text coverage, and lexical properties of fixated words including word length, corpus frequency, part of speech, and GPT-2 surprisal where available. Final trial fixation count, final paragraph reading time, answer choice, correctness, answer/feedback gaze, and all post-horizon fixations were excluded.

As a prospective robustness analysis, we also constructed fixed-time horizons at 2, 4, 6, 8, and 10 s. A fixation contributed only if it had ended by the cutoff, preventing future fixation duration from leaking backward into the prediction. Feature pre-processing was fitted on training data only, including imputation and scaling. The final audit independently recomputed 500 fixation-horizon and 500 time-horizon trial-feature rows directly from the raw fixation archives with zero discrepancies beyond floating-point tolerance.

For each gaze horizon, we compared a text-only model with a text + gaze model using identical trial sets and the same train–test folds. The text representation, classifier, and pre-processing procedure were held constant within each comparison so that differences in predictive performance reflected the incremental contribution of gaze. Pre-processing was estimated from training data only. The same comparison was conducted at 5, 10, 20, and 40 fixations and for the full eye-movement record.

As a stricter test of text separation, we also evaluated reader history under whole-article holdout. When a test trial belonged to a given article, all paragraphs from that article were excluded from the reader's support data. The resulting reader-history estimate therefore relied only on responses to other articles, preventing information from another paragraph of the same article from contributing to prediction of the held-out trial.

2.7. Conditional Gaze Analysis

To test whether current-trial gaze provided predictive information beyond stable reader propensity, we compared the empirical-Bayes reader-history estimate with models that additionally included gaze features from the current trial. Comparisons used the same held-out trials and evaluation folds, with gaze represented at the temporal prefixes described above and using the full eye-movement record. Reader-history estimates were based only on permitted training/support outcomes, and gaze pre-processing was fitted using training data only.

We also conducted a stricter operational test that represented both reader history and text-based item information. Because labelled population item history is not available for a genuinely unseen test text without violating the unseen-text design, text features were used to represent the available item-level information. In ordinary reading, we compared text + reader history with text + reader history + gaze across ten evaluation folds and supplemented the paired fold-level comparison with crossed reader-by-item bootstrap resampling.

To examine whether reader history accounted for the positive gaze increments observed in information seeking, we additionally fitted four models under matched whole-article holdout procedures in both reading conditions: text, text + gaze, text + reader history, and text + reader history + gaze. These comparisons were conducted at 20 fixations and for the full eye-movement record using ten whole-article holdout folds. Within each condition and horizon, the models were evaluated on matched held-out trials. We quantified the incremental contribution of gaze before and after including reader history and tested the

difference between these increments using paired fold-level comparisons, exact two-sided sign-flip tests, and crossed reader-by-item bootstrap resampling. The matched whole-article specification was applied consistently across both reading conditions.

2.8. Released EyeBench Comparisons

For the principal comparison, we used the released trial-level EyeBench MAG-Eye predictions and aligned reader-history predictions for the same 8746 held-out observations. In the seen-reader/unseen-text regime, the benchmark models were trained with labelled responses from permitted training trials, whereas the headline empirical-Bayes reader-history baseline explicitly used training-plus-validation support. The neural models were not jointly retrained with this history prior, so the headline comparison is not a controlled measure of the incremental effect of adding reader history to MAG-Eye. We therefore additionally conducted a matched training-only comparison, restricting held-out evaluation to readers with at least one labelled training response and excluding validation responses from the reader-history estimator.

We determined the training-group assignments using the archived test partitions and OneStop outcomes. Each of ten test folds excludes one disjoint group of 18 readers and one disjoint text group. Excluding the corresponding test groups and the preceding fold's validation groups for readers and texts yielded 144 training readers per fold. These assignments reproduced all ten archived training-support trial counts and training-only fold AUROCs exactly. We estimated beta-binomial priors using training responses alone and compared the models on the 7774 test trials for which training history was available, keeping each test observation and MAG-Eye score unchanged. The fold assignments, training labels, prior parameters, analysis script and matched predictions are provided in the reproducibility archive.

2.9. Repeated-Reading Validation

Repeated reading provides a stringent test of what reader history represents. We predicted comprehension in later repeated-reading trials using reader performance history estimated from first-reading data. To prevent memorisation from masquerading as reader ability, the strict reader-history score excluded the reader's first response to the same paragraph that was being predicted. We compared this general reader score with (a) population item history excluding the focal reader, (b) the reader's own previous correct/incorrect response to the same item, and (c) reader + item history. Crossed reader-by-item bootstrap confidence intervals were used for key repeated-reading contrasts.

2.10. Sample Size and Inferential Sensitivity

This study reanalysed fixed public datasets: participant numbers were therefore determined by the source corpora rather than by an analysis-dependent stopping rule. We used all eligible trials defined by the public benchmark folds and did not add or remove participants in response to statistical outcomes. Because the reader-history question was discovered during exploration, no retrospective power calculation is presented as if it had guided sampling. Inference instead emphasises paired effect sizes, confidence intervals, crossed reader-by-item resampling where appropriate, and exact permutation tests when the small number of folds permits them. This approach makes the precision supported by each dataset explicit, particularly for the four-fold SB-SAT and PoTeC comparisons.

2.11. Statistical Inference and Multiplicity

Paired performance differences were calculated within the released evaluation folds. For OneStop reader history versus MAG-Eye, we report both the paired t-test and exact

two-sided sign-flip test. In SB-SAT and PoTeC, each with four folds, the minimum attainable two-sided exact sign-flip p-value is 0.125. We explored the size of conditional-gaze increments using two one-sided tests (TOST) and symmetric AUROC margins of ± 0.005 and ± 0.010. These correspond to half and one AUROC percentage point, respectively, and provide transparent stringent and more permissive sensitivity bounds against which to judge whether an observed increment is small enough to have limited practical discrimination value for this benchmark. These are pragmatic, post hoc sensitivity margins, not independently validated minimal clinically or educationally important differences and not thresholds fixed before analysis. Any equivalence finding applies only to the selected margin, ten paired folds and specified predictive models, rather than to the absence of gaze information. Temporal hypothesis families used Benjamini–Hochberg correction [27]. Clustered reader-by-item bootstrap resampling was used for key trial-level contrasts. Calibration metrics comprised Brier score, log loss, calibration intercept and slope, and ECE10. We report estimates with uncertainty and avoid interpreting a non-significant superiority test as evidence of no effect.

The planned temporal AUROC analyses used a single Benjamini–Hochberg family of 30 paired fold-level tests, defined as three reader/text generalisation regimes × two reading conditions × five horizons. Secondary balanced-accuracy contrasts were treated as a separate family of 30. We report the unadjusted paired t-test p-values, the original BH-adjusted p-values across 30 AUROC tests, and sensitivity-adjusted values within each five-horizon condition-by-regime group in Supplementary Table S7. To test the progressive-horizon component of RQ1 directly, we computed for each fold (i) the slope of the text-plus-gaze minus text AUROC difference across the ordered five horizons and (ii) the paired difference between the full record and the five-fixation horizon. These fold-level contrasts were tested against zero by two-sided paired t-tests and exact sign-flip tests, with results presented in Supplementary Table S8. Reliability-bin counts for the calibration analyses are provided in Supplementary Table S9.

For the central OneStop comparison between reader history and MAG-Eye, we additionally performed a crossed reader-by-item bootstrap on the aligned trial-level out-of-fold predictions. Reader and item clusters were independently resampled with replacement, and each trial was weighted according to the joint resampling frequency of its reader and item. AUROC differences were recomputed across 2000 bootstrap resamples. This analysis complements the official fold-wise EyeBench comparison by providing uncertainty that accounts for the crossed clustering of observations within readers and items.

3. Results

Results are reported in six subsections covering the temporal contribution of gaze, reader-history performance across datasets, the effect of support size and trial order, calibration and psychometric robustness, incremental gaze effects and repeated reading validation.

The processed fixation dataset contained 2,400,788 fixation observations from all 360 OneStop participants. We identified 24,046 paragraph-reading trials in total and 19,438 non-practice first-reading trials. The first-reading counts matched the released study design exactly: 9718 ordinary-reading and 9720 information-seeking trials. All 19,438 first-reading trial identifiers were unique. Correct responses accounted for 84.06% of first-reading outcomes. Temporal feature matrices contained the same trial rows at each 5, 10, 20, 40, and full horizon within the matched evaluation, so changes across horizons cannot be attributed to horizon-dependent test-set composition.

3.1. Planned Temporal-Gaze Analysis: Small and Non-Monotonic Incremental Effects

The planned analysis evaluated text-only versus text-plus-gaze prediction at five fixation horizons for each reading condition and reader-generalisation regime. In the primary seen-reader/unseen-text setting, ordinary-reading ΔAUROCs were −0.010, −0.010, −0.006, −0.013 and −0.009 for 5, 10, 20, 40 fixations and full reading, respectively (Figure 1). Information-seeking differences were +0.008, +0.023, +0.030, +0.027 and +0.034. The corresponding two-sided unadjusted fold-level p-values and the BH-adjusted values for the 30-test AUROC family are reported in Supplementary Table S7. Under that reported family of 30 AUROC contrasts, the information-seeking 20-fixation and full-record contrasts had adjusted $p = 0.046$ and 0.027, respectively, whereas the 40-fixation contrast had adjusted $p = 0.067$. Under a narrower post hoc five-test family restricted to information seeking in the seen-reader/unseen-text regime, the adjusted p-values would instead be 0.042, 0.040 and 0.045 for 20 fixations, full reading and 40 fixations, respectively. These are tests of an effect at an individual horizon, not tests of an increasing trend across horizons.

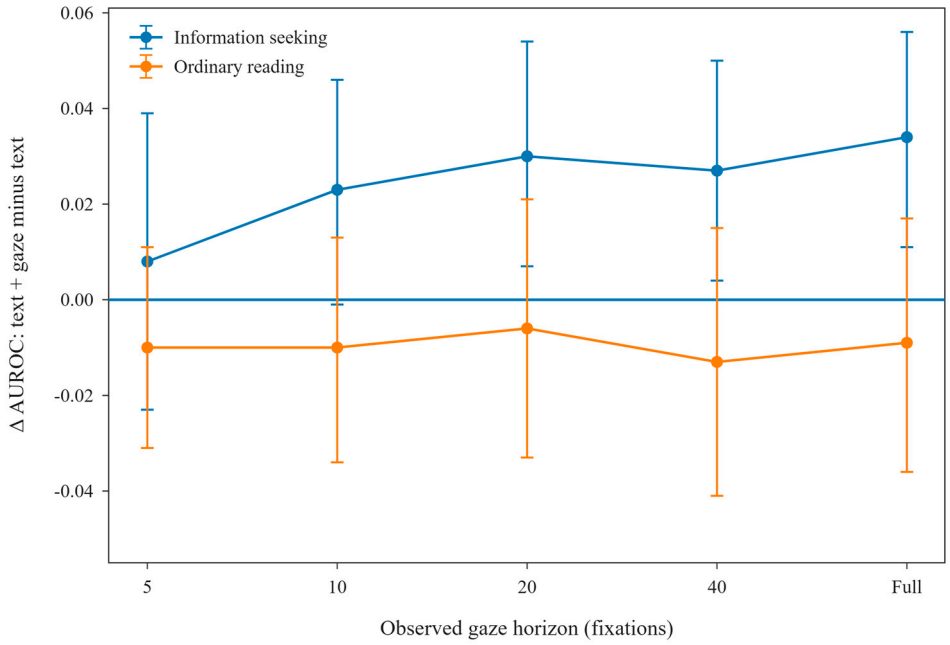

Figure 1. Planned temporal-gaze analysis in the seen-reader/unseen-text regime for ordinary reading and information seeking. Points show fold-level mean AUROC differences between text + gaze and text; error bars show 95% confidence intervals. Incremental gaze effects were small, non-monotonic, and reading-goal dependent.

We therefore tested the increase across horizons directly using the same ten folds. In ordinary reading, the mean linear slope per ordered horizon step was −0.00015 (95% CI [−0.00475, 0.00444], two-sided $p = 0.942$), and the full-record minus five-fixation contrast was +0.00060 (95% CI [−0.01837, 0.01958], $p = 0.944$). For information seeking, the slope was +0.00553 (95% CI [−0.00108, 0.01215], $p = 0.091$), and the full-record minus five-fixation contrast was +0.02566 (95% CI [−0.00295, 0.05427], $p = 0.073$). Exact two-sided sign-flip tests also did not establish a positive trend (Supplementary Table S8). The planned monotonic-increase hypothesis was therefore not established at the two-sided 5% level. The descriptive size and direction of gaze increments also varied among the three generalisation regimes (Supplementary Tables S7 and S8), but these results do not establish a statistically significant regime interaction.

A separate training-only reader fixed-effect analysis, which used sparse text features and one-hot reader indicators rather than the released EyeBench predictors, is documented in Supplementary Table S3.

3.2. Reader-History Performance Across Datasets

In the official EyeBench ordinary-reading, seen-reader/unseen-text folds, the empirical-Bayes reader-history estimator using labelled training-plus-validation support achieved mean AUROC = 0.619, compared with 0.577 for MAG-Eye. The paired fold-mean difference was ΔAUROC = 0.042, 95% CI [0.013, 0.071], with paired *t*-test $p = 0.0096$ and exact two-sided sign-flip $p = 0.0156$ (Table 1). A stricter training-only estimator achieved AUROC = 0.601 against the same 0.577 comparator, with ΔAUROC = 0.024, 95% CI [−0.003, 0.052], paired *t*-test $p = 0.078$. Thus, the stronger headline contrast partly depends on the permitted labelled support set. On the aligned 8746 held-out trials, pooled AUROC was 0.608 for reader history and 0.557 for MAG-Eye (ΔAUROC = 0.051). A crossed reader-by-item bootstrap using 2000 resamples produced a 95% CI [0.002, 0.097] and $p = 0.042$. This pooled difference is not identical to the equally weighted fold-mean comparison because the two calculations weight the data differently. These results support reader performance history as a relevant comparator in personalised comprehension prediction, but they do not establish superiority under all support rules or jointly trained architectures.

Table 1. Reader performance history versus the strongest released unseen-text benchmark.

Dataset	Folds	Reader-History AUROC	Comparator (AUROC)	Paired ΔAUROC [95% CI]	Exact Sign-Flip *p*
OneStop	10	0.619	MAG-Eye (0.577)	+0.042 [0.013, 0.071]	0.016
SB-SAT	4	0.559	PostFusion-Eye (0.559)	+0.001 [−0.151, 0.153]	1.000
PoTeC	4	0.611	Logistic regression (0.559)	+0.052 [−0.024, 0.128]	0.250

A matched training-only analysis retained 7774 held-out trials from 144 readers per fold with at least one training response (Supplementary Table S10). Across the ten folds, training-only reader history achieved mean AUROC = 0.6101 and MAG-Eye mean AUROC = 0.5785 on those same trials (mean paired ΔAUROC = +0.0316, 95% CI [0.0010, 0.0622]; paired *t*-test $p = 0.0444$; exact two-sided sign-flip $p = 0.0508$). The estimated difference favoured reader history, although the exact sign-flip test did not reach the 0.05 significance threshold. The fold assignments were verified against archived support totals and earlier training-only AUROCs. These results support the use of reader history as a comparator without establishing universal superiority or the performance of jointly trained models. Fold-specific beta-prior parameters for the separate training-plus-validation predictions are reported in Supplementary Table S11.

The SB-SAT and PoTeC results were treated as exploratory cross-dataset consistency checks because each dataset provided only four benchmark folds (Table 1). In SB-SAT, reader history achieved AUROC = 0.559 and PostFusion-Eye achieved 0.559 (paired ΔAUROC = 0.001, 95% CI [−0.151, 0.153]). In PoTeC, reader history achieved 0.611 compared with 0.559 for the strongest released comparator, logistic regression (paired ΔAUROC = 0.052, 95% CI [−0.024, 0.128]; exact sign-flip $p = 0.250$). The intervals are too wide to establish superiority or equivalence in either dataset. In particular, the SB-SAT interval permits substantial effects in both directions. These descriptive patterns do not amount to inferential replication or demonstrate broad generalisability.

3.3. Few-Shot and Prospective Reader History

Reader performance history improved systematically as labelled support increased. In OneStop, averaging across 30 independent support draws per fold, AUROC rose from 0.521 with one response to 0.542 with three, 0.556 with five, 0.575 with ten, 0.595 with

twenty, 0.613 with forty, and 0.619 with all available support. Relative to the MAG-Eye fold mean of 0.577, one to five responses were significantly worse, ten responses produced a point estimate close to the comparator (ΔAUROC = −0.001, 95% CI [−0.021, 0.019]), but equivalence was not established, twenty were modestly higher but uncertain, and forty were reliably higher (ΔAUROC = 0.037, 95% CI [0.009, 0.064], exact $p = 0.020$). The aim is not to claim equivalence between ten questions and an eye tracker, but to show how rapidly a simple behavioural history enters the same predictive performance range (Figure 2b; Supplementary Table S4).

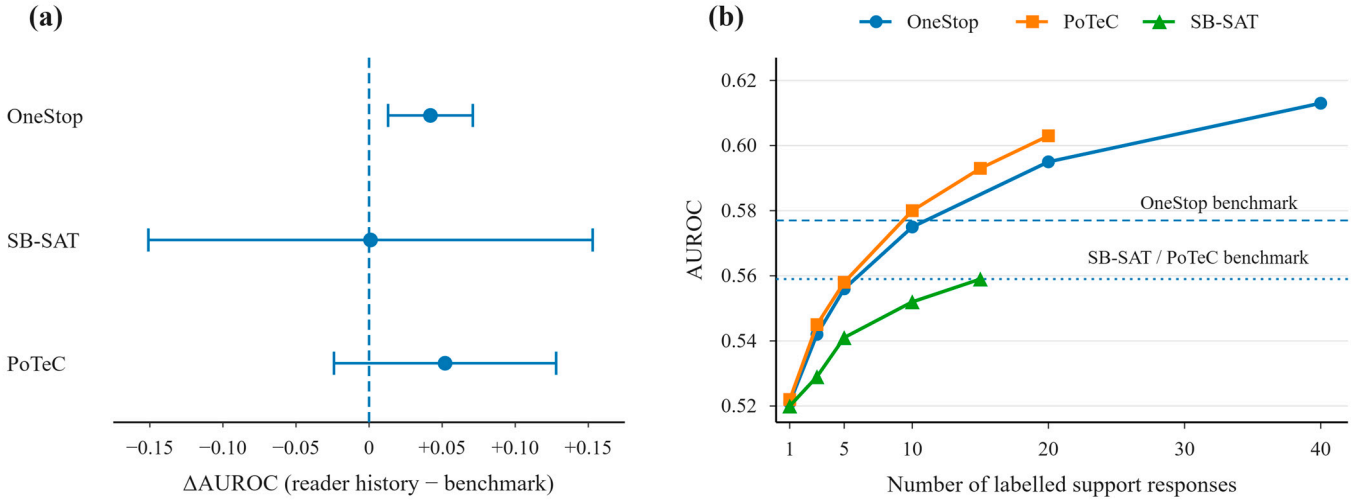

Figure 2. Reader-history performance across datasets. (**a**) Paired AUROC difference against the strongest released benchmarks (95% confidence intervals). (**b**) AUROC by number of labelled support responses where horizontal lines indicate benchmark performance.

A broadly increasing descriptive dose-response pattern was also observed in the two external datasets. PoTeC increased from approximately 0.522 with one support answer to 0.558 with five, 0.580 with ten, 0.603 with twenty, and 0.611 with all support. SB-SAT increased from approximately 0.520 with one to 0.541 with five and 0.559 with fifteen/all support. These smaller datasets cannot identify a precise support-size threshold, and the descriptive patterns require confirmation with larger external samples.

The OneStop prospective analysis restricted reader history to responses that occurred earlier in the session, as determined from the original trial indices and timestamps. Mean AUROC increased from 0.523 with one earlier answer to 0.545 with three, 0.555 with five, 0.575 with ten, 0.591 with twenty, and 0.599 using all available earlier support. Approximately 90.1% of evaluated cases had at least one earlier history response and 88.0% had at least five; these support-coverage figures are noted alongside the prospective results in Supplementary Table S6. Although prospective performance was lower than the unconstrained support protocol, the monotonic trend demonstrates that the reader signal is operationally available during an ongoing session and is not dependent on using labels that occur later in time.

3.4. Calibration and Psychometric Robustness

Empirical-Bayes shrinkage improved probability estimates relative to raw reader accuracy and a fold-specific constant-probability predictor. In ordinary reading, the empirical-Bayes estimator achieved Brier = 0.1491, log loss = 0.4718, ECE10 = 0.0034 and calibration slope = 0.9621, compared with raw-history Brier = 0.1501 and constant-predictor Brier = 0.1529. In information seeking, the corresponding Brier scores were 0.1119 (empirical Bayes), 0.1128 (raw history) and 0.1137 (constant predictor); empirical-Bayes ECE10 was

0.0050 and calibration slope was 0.9408. For raw reader history, ECE10 was 0.0245 and the calibration slope was 0.5675 in ordinary reading, the corresponding values in information seeking were 0.0231 and 0.3632. These values indicate overconfidence in raw reader accuracy. These comparisons describe probability quality under the observed reading regimes and sampling conditions. The near-unit slopes and low calibration errors are consistent with good calibration in the present data, but do not establish that calibration or predictive accuracy will generalise to a new dataset (Figure 3; Supplementary Tables S5 and S9).

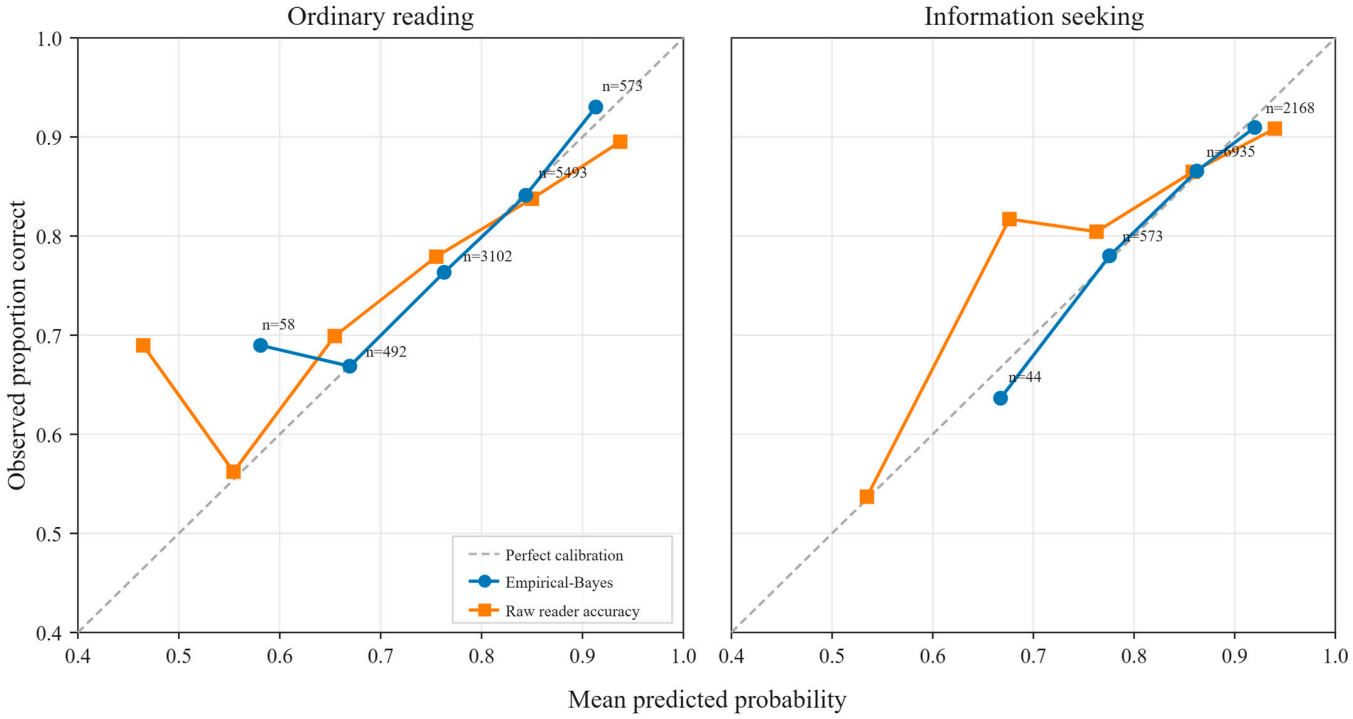

Figure 3. Reliability curves for empirical-Bayes and raw reader-history estimates in OneStop ordinary reading and information seeking. The labelled points show the number of trials in each non-empty empirical-Bayes calibration bin.

The reader signal survived stricter text separation. In canonical whole-article holdout, in which all paragraphs from each held-out article were excluded from support, reader performance history achieved mean AUROC = 0.605 (SD = 0.019) in ordinary reading and 0.602 (SD = 0.029) in information seeking. These figures come from a separate whole-article holdout robustness analysis and are not interchangeable with the official EyeBench fold-mean benchmark value. Because the held-out articles are entirely unseen, the result cannot be explained by support from another paragraph within the same article.

Regularised reader + item psychometric models produced a nearly identical reader-only signal after adjusting for support-item difficulty. Across regularisation values spanning C = 0.01 to 100, reader AUROC remained approximately 0.615–0.618 in OneStop, 0.610–0.611 in PoTeC, and 0.559 in SB-SAT. The persistence of the reader term after item adjustment argues against the explanation that high-performing readers simply encountered easier support questions.

Reader and population item histories each predicted comprehension, and their combination achieved a higher AUROC. In ordinary OneStop first-reading trials, leave-one-out reader history achieved AUROC = 0.612, population item history 0.642, and reader + item history 0.678. In information seeking, the corresponding values were 0.602, 0.635, and 0.673. The decomposition supports a psychometric interpretation in which reader and item histories provide complementary predictive information. It also sharpens the role of

gaze: a process measure should ideally explain the remaining reader-by-item interaction or transient state after these stable components are accounted for. The complementary effects of reader and item history are also illustrated in Supplementary Figure S2.

3.5. Incremental Contribution of Current-Trial Gaze

Conditional gaze analyses tested whether the current eye-movement provided additional predictive information after accounting for reader performance history. In ordinary reading, empirical-Bayes reader history alone achieved AUROC = 0.612. Adding full-record gaze increased AUROC to 0.615. In information seeking, reader history alone achieved AUROC = 0.604, compared with 0.595 after adding full record gaze. Across the evaluated fixation horizons, the tested gaze features did not provide consistent incremental improvements over reading history.

A stricter ordinary-reading analysis represented both text information and reader history. The text + reader-history model achieved mean fold AUROC = 0.6159, compared with 0.6173 after including gaze (ΔAUROC = +0.00134, 95% CI [−0.00200, 0.00469], $p = 0.388$). Exploratory paired-fold TOST results supported equivalence at the post hoc margins of ± 0.005 ($p = 0.0177$) and ± 0.010 ($p < 0.001$). The corresponding crossed reader-by-item bootstrap yielded ΔAUROC = +0.0009, 95% CI [−0.0044, 0.0067], $p = 0.751$, across 1500 resamples. This interval was contained within ± 0.010 but not ± 0.005. Equivalence at the tighter margin was therefore not robust to reader-by-item resampling. The reader-history-only comparisons did not establish equivalence at either margin in ordinary reading (TOST $p = 0.3552$ and 0.0812) or information seeking ($p = 0.6610$ and 0.4665).

We further examined whether reader history accounted for the positive gaze increments observed in information seeking in the planned temporal-gaze analysis (Section 3.1). Using matched whole-article holdout procedures in both reading conditions, we compared the incremental contribution of gaze before and after including reader history (Table 2; Supplementary Table S12).

Table 2. Incremental predictive contribution of current-trial gaze representations after accounting for reader history and available text information.

Analysis	Horizon	Baseline AUROC	With Gaze AUROC	ΔAUROC
Reader-history baseline				
Ordinary, reader history	Full	0.612	0.615	+0.003
Information seeking, reader history	Full	0.604	0.595	−0.009
Ordinary, text + reader history	Full	0.616	0.617	+0.001
Matched whole-article holdout: text + reader history				
Ordinary reading	20	0.620	0.613	−0.007
Ordinary reading	Full	0.620	0.610	−0.010
Information seeking	20	0.583	0.578	−0.005
Information seeking	Full	0.583	0.579	−0.004

Note: Values are mean fold-level AUROCs. Differences represent the incremental contribution of gaze relative to the corresponding baseline. Confidence intervals and exploratory equivalence tests are reported in Section 3.5. Full matched whole-article holdout comparisons are reported in Supplementary Table S12. Bolds are section headings within the table and distinguish two groups of analyses conducted under different evaluation procedures.

In information seeking, adding gaze to text increased AUROC by 0.0192 at 20 fixations and 0.0298 for the full record. When reader history was included, the corresponding gaze increments were −0.0051 and −0.0041. The reduction in incremental predictive performance was −0.0243 at 20 fixations (95% paired CI [−0.0414, −0.0072], exact sign-flip $p = 0.0156$) and −0.0339 for the full record (95% paired CI [−0.0493, −0.0185], exact sign-flip $p = 0.0059$). Crossed reader-by-item bootstrap intervals supported both reductions (20 fixations: 95% CI [−0.0463, −0.0040]; full record: 95% CI [−0.0591, −0.0120]).

In ordinary reading, the corresponding changes in the incremental contribution of gaze were +0.0022 at 20 fixations (exact sign-flip $p = 0.8457$) and +0.0002 for the full record ($p = 0.9844$). Thus, the information-seeking results showed a reduction in the incremental contribution of gaze after including reader history, whereas ordinary reading showed no corresponding reduction in a condition where gaze had already provided little predictive improvement over text. The information-seeking full-record conditional-gaze contrast yielded a crossed-bootstrap 95% CI of [−0.0196, 0.0115]. Exploratory equivalence tests did not establish equivalence at either ±0.005 ($p = 0.4085$) or ±0.010 ($p = 0.0711$). The attenuation analyses were exploratory and their p-values were not adjusted for multiplicity.

These results suggest that part of the predictive contribution of gaze over text in personalised information-seeking prediction overlaps with information already available from reader performance history. They do not establish the general absence of trial-specific gaze information.

3.6. Repeated-Reading Validation

Reader history learned from first-reading data generalised to later repeated-reading comprehension even after the focal reader's previous answer to the same paragraph was removed from the history. In ordinary repeated reading, strict reader history achieved AUROC = 0.631, population item history 0.650, the reader's same-item previous answer only 0.543, and reader + item history 0.696. In information seeking, the corresponding values were 0.608, 0.647, 0.507, and 0.679 (Table 3; Supplementary Figure S3). Crossed bootstrap resampling confirmed that reader history outperformed the same-item previous response in both ordinary reading (ΔAUROC = 0.087, 95% CI [0.019, 0.151], $p = 0.018$) and information seeking (ΔAUROC = 0.101, 95% CI [0.010, 0.183], $p = 0.034$). Adding item history to reader history also improved both conditions. General performance across other texts was therefore more informative than remembering whether the reader had previously answered this exact paragraph correctly.

Table 3. Strict repeated-reading validation.

Condition	Reader History	Item History	Same-Item Previous Answer	Reader + Item
Ordinary	0.631	0.650	0.543	0.696
Information seeking	0.608	0.647	0.507	0.679

Across analyses, reader-history performance was examined under (a) strict training-only estimation, (b) a matched training-only comparison restricted to test readers with observed training labels, (c) whole-article holdout, (d) empirical-Bayes shrinkage and probability calibration, (e) regularised reader-item adjustment, (f) few-shot and prospective chronological support, (g) strict repeated-reading validation and (h) exploratory external evaluation in SB-SAT and PoTeC. The matched comparison showed a positive estimated AUROC difference against MAG-Eye, with an exact sign-flip p-value slightly above 0.05. No individual analysis establishes universal superiority over gaze-based prediction, but these findings justify treating reader-history baselines as an important control in personalised comprehension studies.

4. Discussion

The present study examined the contribution of current-trial gaze to personalised reading-comprehension prediction and whether stable reader performance history provides an important baseline for interpreting that contribution. The planned temporal-gaze analysis addressed RQ1 and did not support the hypothesised monotonic increase in pre-

dictive value as progressively more of the fixation sequence became available. In the seen-reader/unseen-text regime, gaze did not improve prediction over text in ordinary reading, whereas information seeking showed positive text-adjusted increments at 20 fixations and for the full record (BH-adjusted $p = 0.046$ and 0.027 within the planned 30-test family). Differences among regimes and reading conditions should not be interpreted as a tested interaction (Supplementary Table S7).

The exploratory reader-history analyses addressed RQ2 and RQ3. In the personalised seen-reader/unseen-text setting, labelled outcomes from the same reader formed a useful baseline for trial-level prediction. The primary training-plus-validation benchmark difference favoured reader history, while the matched training-only analysis found a smaller positive mean difference (ΔAUROC = 0.0316; paired t-test $p = 0.0444$; exact sign-flip $p = 0.0508$). This qualifies rather than negates the methodological importance of reader-history controls. The external estimates from SB-SAT and PoTeC remain too uncertain for inferential replication claims.

RQ3 examined whether current-trial gaze provided additional predictive information once reader history and available text-based item information were represented. Reader and item histories provided complementary predictive information. In ordinary reading, adding gaze to text and reader history increased mean AUROC from 0.6159 to 0.6173. Exploratory paired-fold equivalence tests supported both post hoc margins of $\pm$0.005 and $\pm$0.010, although the crossed reader-by-item bootstrap interval was contained only within the wider $\pm$0.010 margin. Equivalence at the tighter margin was therefore not robust to reader-by-item resampling.

In information seeking, gaze improved over text at 20 fixations and for the full record in the planned analysis. However, the positive gaze increments were reduced after including reader history, with these reductions supported by exact paired tests and crossed reader-by-item bootstrap intervals at both horizons. This suggests that reader history accounts for part of the positive information-seeking gaze increment. The possibility that gaze partly reflects stable reader differences is therefore consistent with the findings and previous evidence [12,16], although this does not establish a definitive mechanism. Future models should jointly control for text and reader history in both reading conditions before gains from gaze are interpreted as trial-specific cognitive information.

4.1. A Psychometric Interpretation of Multimodal Comprehension Prediction

The results align naturally with a person–item decomposition. Let θp denote reader propensity, βi item difficulty, and γ_{pi} the residual reader-by-item interaction. Reader performance history estimates θp, item history and text models estimate βi and current-trial gaze is most scientifically valuable if it helps explain γ_{pi} or transient state variation. In our OneStop decomposition, reader + item history reached AUROC near 0.68, showing complementary predictive contributions but not quantifying the amount of residual outcome variance. This changes the interpretation of multimodal improvements. A gain over text alone does not necessarily demonstrate that gaze captures moment-to-moment comprehension difficulty; it may partly recover stable reader identity or ability that a text model cannot know.

This interpretation is compatible with, rather than opposed to, the eye–mind link. Eye movements can reflect online processing while simultaneously containing strong stable between-reader structure. The two sources are statistically entangled. A useful future model should therefore aim explicitly at the conditional quantity $I\,(C; G \mid \text{reader, item})$, not merely $I\,(C; G \mid \text{text})$. Such a model would test whether gaze identifies occasions when a normally successful reader is unexpectedly struggling with an otherwise manageable item, or when a generally weaker reader succeeds through unusually effective processing.

4.2. Implications for EyeBench and Future Benchmarks

EyeBench makes an important methodological contribution by standardising tasks, datasets, models, and generalisation regimes. The present results suggest one extension to the benchmark hierarchy. In addition to majority, reading-speed, text-only, traditional gaze, and multimodal baselines, the seen-reader/unseen-text regime should include a labelled reader-history baseline. For datasets with repeated questions or sufficient trials, a reader + item history baseline is also informative. New gaze models could then be asked to demonstrate incremental performance beyond this stable structure. This would make benchmark gains more interpretable and would reduce the risk of rewarding models that primarily learn reader identity.

The proposal does not apply unchanged to every deployment setting. For a completely unseen reader, labelled history is unavailable and gaze may be valuable precisely because it can estimate reader characteristics implicitly. The strongest design may therefore be a two-stage system: population-level inference at cold start, followed by explicit behavioural reader history as labelled outcomes accumulate. The prospective OneStop curve provides an empirical starting point for this idea, showing that reader history becomes useful within a relatively small number of earlier questions.

4.3. Implications for Adaptive Reading and Educational Technology

Adaptive reading systems need to decide not only whether a prediction is accurate, but also which information source is worth collecting. EyeLiveMetrics demonstrates the technical feasibility of computing word- and paragraph-level gaze measures in real time and using them immediately in an interactive interface [28]. Such systems could, in principle, adapt text difficulty, offer clarification, or defer intervention when comprehension appears secure. Related interactive-reading work has also shown the feasibility of combining digital text presentation with dedicated and webcam-based eye-tracking, while directly comparing gaze-based and non-gaze interaction modes [29]. Our findings motivate future tests of systems that combine these process signals with a reader-history prior. Whether such integrated models improve prediction or intervention decisions remains unresolved.

The cost comparison is also relevant as eye-tracking moves outside specialised laboratories. High-precision infrared systems remain the benchmark for fine-grained reading research, but access is constrained by equipment costs and controlled testing requirements. Webcam eye-tracking has improved markedly and can reproduce canonical reading effects, yet recent co-registration work still finds substantially more spatial noise and poorer online quality than EyeLink measurements [30]. Reader performance history is therefore attractive as a low-cost complement. A practical implication is to evaluate gaze alongside lower-cost behavioural information rather than assuming that gaze must provide unique information in every setting.

OneStop's ordinary and information-seeking regimes, together with recent evidence that repeated reading leaves a decodable gaze signature [31], also clarify why generic gaze-to-comprehension models may be unstable. Reading goals alter eye movements rapidly and can themselves be decoded shortly after reading begins [23]. In our planned temporal analysis, information seeking showed positive gaze increments over text at two horizons for known readers on unseen texts, whereas ordinary reading did not. The matched whole-article holdout analyses further showed that these positive gaze increments were reduced after accounting for reader history and text information. The predictive meaning of regressions, long fixations or coverage patterns may therefore depend on the reading task. This contextual dependence is consistent with previous evidence that eye-movement patterns vary with both the reading medium and the linguistic demands of the material [10].

A single gaze signature is unlikely to capture every reading task and goal. Models that use gaze diagnostically should condition on reading goal, task structure, and reader baseline.

Probability calibration is often neglected in cognitive prediction work, yet it is central to adaptive decisions. A system that intervenes when predicted comprehension falls below a threshold implicitly treats model probabilities as meaningful quantities. The empirical-Bayes reader estimator produced near-unit calibration slopes and very small ECE10 in both OneStop regimes, while raw reader accuracy was markedly overconfident. This illustrates a broader point: strong AUROC alone is insufficient for personalised applications. Calibration should accompany discrimination in future eye-tracking benchmarks, especially if model outputs are used to trigger educational or accessibility interventions.

This work also illustrates the scientific value of reusable eye-tracking corpora. The reader-history result was not foreseeable from a single summary table; it required trial-level labels, reader identifiers, benchmark fold definitions, released model outputs, fixation reports, repeated-reading structure, and metadata. The growing ecosystem of eye-tracking corpora will support such reuse when data formats, preprocessing parameters, metadata, and code are documented consistently [32]. The FAIR principles extend this requirement to analytical workflows themselves [33].

4.4. Limitations

These findings require several qualifications. First, reader performance history requires prior labelled responses and cannot provide personalised predictions for completely new readers without an onboarding phase. Second, comparison of separately trained MAG-Eye and reader-history models does not establish the benefit of a jointly trained reader-text-gaze model. Third, the matched training-only comparison used fold assignments determined from the archived held-out partitions and verified against earlier fold-level AUROCs and training-support counts. Independent confirmation against the original official EyeBench training and validation manifest remains unavailable. The estimated difference favoured reader history, although the exact two-sided sign-flip p-value was slightly above 0.05. Fourth, the 972 ordinary-reading trials without paired MAG-Eye outputs have now been assigned exactly to the unseen-reader/unseen-text intersection of the archived benchmark folds.

Fifth, the external datasets each have four benchmark folds. Their wide confidence intervals and limited inferential resolution do not support claims of inferential replication, superiority or equivalence in SB-SAT or PoTeC. Sixth, reader history may summarise multiple constructs: language skill, comprehension strategy, motivation, fatigue propensity, test-taking skill, and domain knowledge. Available OneStop metadata did not reveal a single trivial explanation, and item adjustment did not remove the signal, but the latent construct should not be labelled 'ability' too narrowly without dedicated measurement. Seventh, although the matched whole-article holdout analyses showed a reduction in the incremental predictive contribution of gaze after including reader history in information seeking, these analyses were exploratory and used a specific set of gaze representations and model specifications. The results do not establish a general absence of trial-specific gaze information, a causal explanation of the observed attenuation, or the performance of jointly trained sequence-based models. Finally, the central reader-history findings arose from exploratory analysis and were followed by internal robustness checks and limited cross-dataset evaluations; they are not independent preregistered confirmation. External prospective studies designed around reader performance history would provide stronger confirmatory evidence.

4.5. Future Work

A next modelling step is to integrate reader history directly into multimodal architectures and retrain end-to-end under identical benchmark folds. The reader-history posterior could be represented as a scalar prior, a learned reader token, or a low-dimensional embedding whose uncertainty depends on support size. Hierarchical models could jointly estimate reader, item, and gaze-interaction components so that uncertainty for sparsely observed readers propagates naturally. Adaptive item selection could then ask which labelled responses are most informative for constructing the reader-history prior.

The second direction is to move from outcome prediction to residual diagnosis. Instead of predicting raw correctness, gaze models could predict departures from an expected reader-item probability. This target is closer to the practical question an adaptive system needs to answer: is this reader showing a lower probability of answering the current item correctly than would be expected from their previous performance and the difficulty of the item? Such residualised targets may reduce reader-identity shortcuts and make moment-to-moment gaze information more visible. Repeated-reading and information-seeking manipulations offer useful test beds because they alter processing while holding substantial reader and text structure constant.

Replication should also extend beyond high-precision laboratory eye-tracking. The growing feasibility of webcam-based reading studies creates an opportunity to test whether explicit reader-history baselines become even more important as gaze becomes noisier. Recent webcam-based reading work suggests that canonical reading effects can be recovered despite substantially greater spatial noise and lower measurement quality than laboratory EyeLink recordings [30]. In such settings, labelled comprehension responses collected from the reader on previous trials provide a comparatively low-cost source of stable reader information that does not depend on precise gaze measurement. This information could account for stable between-reader differences first, allowing lower-precision gaze measures to be evaluated for the dynamic information they may uniquely provide, such as attention loss, re-reading or local processing difficulty.

5. Conclusions

The present study examined whether current-trial gaze improves personalised reading-comprehension prediction after reader history and text information are considered. Direct fold-wise tests across five ordered gaze horizons did not establish the hypothesised monotonic increase in gaze-related predictive value. Subsequent exploratory analyses showed that labelled reader performance history provides a baseline when previous comprehension responses from the same reader are available. In OneStop, training-plus-validation reader history exceeded the strongest released benchmark on average. A matched training-only comparison across 7774 test trials showed mean AUROC 0.6101 for reader history versus 0.5785 for MAG-Eye. Reader and item histories provided complementary predictive information.

The matched whole-article holdout analyses showed that the positive predictive contribution of gaze over text in information seeking was reduced after including reader history, both at 20 fixations and for the full eye-movement record. These findings suggest that part of the predictive contribution of gaze in personalised comprehension prediction reflects stable differences between readers.

We therefore recommend that personalised reading-comprehension benchmarks include explicit reader-history baselines when evaluating gaze-based models. The present findings do not establish that eye movements are generally uninformative or that reader-history models outperform all multimodal alternatives. Future research should evaluate jointly trained reader, text and gaze models under prospectively specified evaluation protocols.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/jemr19060133/s1. Figure S1: Conceptual decomposition of personalised comprehension prediction, whereby trial-level comprehension is computed by considering the interplay of stable components (i.e., reader-level and item-level) and dynamic components (i.e., trial-specific). Figure S2: Reader and item histories provide complementary predictive information in OneStop first reading. The combined model substantially outperforms either stable component alone. Figure S3: Repeated-reading validation. General reader history predicts later repeated-reading comprehension more strongly than the reader's previous answer to the same item; reader and item histories are complementary. Table S1: Summary of the roles of the datasets included in the analysis. Table S2: Analyses, design and evidence status of each variable considered in the study. Table S3: Training-only personalised baseline in OneStop ordinary reading. Table S4: Few-shot reader performance history. Table S5: Probability calibration of reader-history estimators in OneStop. Table S6: Prospective OneStop reader history using only genuinely earlier responses. Table S7: Full temporal AUROC contrasts for three generalisation regimes, two reading conditions and five gaze horizons. The original BH correction used all 30 AUROC contrasts as one family. The five-test BH values are provided for sensitivity only. All p-values are two-sided fold-level paired t-tests. Table S8: Direct tests of the increasing-horizon component of RQ1. Each row uses ten paired fold-level ΔAUROC values; the slope uses the five ordinal horizon positions (5, 10, 20, 40, full), not fixation-count units. Table S9: Reliability-bin counts for Figure 3. The full numerical table is supplied as calibration_bins_with_counts.csv in the accompanying reproducibility archive. The lowest empirical-Bayes ordinary-reading bin contains 58 observations; mean prediction 0.581 and observed accuracy 0.690. Table S10: Matched training-only reader-history versus MAG-Eye comparison using the OneStop seen-reader/unseen-text folds. Both methods are evaluated on the same 7,774 held-out test trials from readers with at least one training label; no validation outcomes are used to estimate reader history. Table S11: Fold-specific beta-prior parameters for the archived training-plus-validation reader-history predictions. The parameters were solved retrospectively from the archived held-out prediction probabilities and the original first-reading response counts, not taken from an original optimiser log. Table S12: Matched whole-article holdout comparisons of the incremental predictive contribution of gaze before and after including reader performance history in OneStop ordinary reading and information seeking. All values represent equally weighted means across ten held-out folds. Models were evaluated on matched held-out trials within each comparison. The accompanying reproducibility archive, including analysis code, derived data and supporting results, is available at https://doi.org/10.5281/zenodo.24117082.

Author Contributions: Conceptualisation, methodology, formal analysis and data curation, A.M.; writing, original draft preparation, A.M. and K.C.; writing, review and editing, K.C.; visualisation, K.C. All authors have read and agreed to the published version of the manuscript.

Funding: Open Access Funding by the University of Vienna.

Institutional Review Board Statement: This study is a secondary analysis of publicly released research datasets. No new participants were recruited, and no new data were collected. Ethics approval for the original data collections is described in the respective source publications. No separate ethics review was conducted for the present secondary analysis.

Informed Consent Statement: Not applicable to new data collection in this study, as no new participants were recruited and only previously collected datasets were reanalysed. The original studies' consent procedures are described in their respective source publications.

Data Availability Statement: The original OneStop dataset is available at https://doi.org/10.17605/OSF.IO/2PRDQ [22]. The original SB-SAT and PoTeC datasets and the EyeBench benchmark are available through the repositories cited in [16,18,26]. The analysis code and derived data supporting the present study are publicly available through Zenodo at https://doi.org/10.5281/zenodo.24117082. The accompanying reproducibility archive includes aligned trial-level predictions, reader and item identifiers, evaluation-fold assignments, empirical-Bayes prior parameters, session-order data, fold-level statistical results and the procedures used to verify the reported findings.

Acknowledgments: We thank colleagues at the University of Vienna for their support. During manuscript preparation, OpenAI ChatGPT (GPT-5.6 Sol, August 2026) assisted with analysis-code development, exploratory analysis planning and language editing. Analysis and figure outputs were checked against saved numerical results. The authors reviewed and verified the material used in the manuscript and take responsibility for the final content and interpretations.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Snow, C.E. *Reading for Understanding: Toward an R&D Program in Reading Comprehension*; RAND Corporation: Santa Monica, CA, USA, 2002. Available online: https://www.rand.org/pubs/monograph_reports/MR1465.html (accessed on 18 November 2026).
2. Kintsch, W. *Comprehension: A Paradigm for Cognition*; Cambridge University Press: Cambridge, UK, 1998.
3. Perfetti, C.; Stafura, J. Word knowledge in a theory of reading comprehension. *Sci. Stud. Read.* **2014**, *18*, 22–37. [CrossRef]
4. Rayner, K. Eye movements in reading and information processing: 20 years of research. *Psychol. Bull.* **1998**, *124*, 372–422. [CrossRef] [PubMed]
5. Just, M.A.; Carpenter, P.A. A theory of reading: From eye fixations to comprehension. *Psychol. Rev.* **1980**, *87*, 329–354. [CrossRef]
6. Clifton, C., Jr.; Ferreira, F.; Henderson, J.M.; Inhoff, A.W.; Liversedge, S.P.; Reichle, E.D.; Schotter, E.R. Eye movements in reading and information processing: Keith Rayner's 40 year legacy. *J. Mem. Lang.* **2016**, *86*, 1–19. [CrossRef]
7. Rayner, K. Eye movements and attention in reading, scene perception, and visual search. *Q. J. Exp. Psychol.* **2009**, *62*, 1457–1506. [CrossRef] [PubMed]
8. Reichle, E.D.; Rayner, K.; Pollatsek, A. The E-Z Reader model of eye-movement control in reading: Comparisons to other models. *Behav. Brain Sci.* **2003**, *26*, 445–476. [CrossRef] [PubMed]
9. Engbert, R.; Nuthmann, A.; Richter, E.M.; Kliegl, R. SWIFT: A dynamical model of saccade generation during reading. *Psychol. Rev.* **2005**, *112*, 777–813. [CrossRef] [PubMed]
10. Mendhakar, A.; Sierak, K.; Chana, K.; Leder, H. Does the reading platform matter? An eye-tracking study. *AILA Rev.* **2025**, *38*, 105–128. [CrossRef]
11. Makowski, S.; Jäger, L.A.; Abdelwahab, A.; Landwehr, N.; Scheffer, T. A discriminative model for identifying readers and assessing text comprehension from eye movements. In *Machine Learning and Knowledge Discovery in Databases, Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2018), Dublin, Ireland, 10–14 September 2018*; Berlingerio, M., Bonchi, F., Gärtner, T., Hurley, N., Ifrim, G., Eds.; Springer: Cham, Switzerland, 2019; Volume 11051, pp. 209–225. [CrossRef]
12. Kuperman, V.; Van Dyke, J.A. Effects of individual differences in verbal skills on eye-movement patterns during sentence reading. *J. Mem. Lang.* **2011**, *65*, 42–73. [CrossRef] [PubMed]
13. Southwell, R.; Gregg, J.; Bixler, R.; D'Mello, S.K. What eye movements reveal about later comprehension of long connected texts. *Cogn. Sci.* **2020**, *44*, e12905. [CrossRef] [PubMed]
14. Mézière, D.C.; Yu, L.; Reichle, E.D.; von der Malsburg, T.; McArthur, G. Using eye-tracking measures to predict reading comprehension. *Read. Res. Q.* **2023**, *58*, 425–449. [CrossRef]
15. Yarkoni, T.; Westfall, J. Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspect. Psychol. Sci.* **2017**, *12*, 1100–1122. [CrossRef] [PubMed]
16. Ahn, S.; Kelton, C.; Balasubramanian, A.; Zelinsky, G. Towards predicting reading comprehension from gaze behavior. In *Proceedings of the 2020 Symposium on Eye Tracking Research and Applications (ETRA 2020), Stuttgart, Germany (Held Virtually), 2–5 June 2020*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 1–5. [CrossRef]
17. Shubi, O.; Meiri, Y.; Hadar, C.A.; Berzak, Y. Fine-grained prediction of reading comprehension from eye movements. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024), Miami, FL, USA, 12–16 November 2024*; Association for Computational Linguistics: Miami, FL, USA, 2024; pp. 3372–3391. [CrossRef]
18. Shubi, O.; Reich, D.R.; Gruteke Klein, K.; Angel, Y.; Prasse, P.; Jäger, L.A.; Berzak, Y. EyeBench: Predictive modeling from eye movements in reading. In Proceedings of the Thirty-Ninth Annual Conference on Neural Information Processing Systems (NeurIPS 2025), Datasets and Benchmarks Track, San Diego, CA, USA, 2–7 December 2025. [CrossRef]
19. Scozzaro, C.J.; Delsanto, M.; Radicioni, D.P. Beyond the average reader: The reader embedding approach. In *Findings of the Association for Computational Linguistics: ACL 2025, Vienna, Austria, 27 July–1 August 2025*; Association for Computational Linguistics: Vienna, Austria, 2025; pp. 15231–15244. [CrossRef]
20. Rasch, G. *Probabilistic Models for Some Intelligence and Attainment Tests*, Expanded ed.; University of Chicago Press: Chicago, IL, USA, 1980.
21. Berzak, Y.; Malmaud, J.; Shubi, O.; Meiri, Y.; Lion, E.; Levy, R. OneStop: A 360-participant English eye tracking dataset with different reading regimes. *Sci. Data* **2025**, *12*, 1995. [CrossRef] [PubMed]

22. Berzak, Y.; Malmaud, J.; Shubi, O.; Meiri, Y.; Lion, E.; Levy, R. *OneStop Eye Movements [Data Set]*; Open Science Framework: Charlottesville, VA, USA, 2025. [CrossRef]
23. Shubi, O.; Hadar, C.A.; Berzak, Y. Decoding reading goals from eye movements. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, Vienna, Austria, 27 July–1 August 2025*; Association for Computational Linguistics: Vienna, Austria, 2025; pp. 5616–5637. [CrossRef]
24. Dunn, M.J.; Alexander, R.G.; Amiebenomo, O.M.; Arblaster, G.; Atan, D.; Erichsen, J.T.; Ettinger, U.; Giardini, M.E.; Gilchrist, I.D.; Hamilton, R.; et al. Minimal reporting guideline for research involving eye tracking (2023 edition). *Behav. Res. Methods* **2024**, *56*, 4351–4357. [CrossRef] [PubMed]
25. Hessels, R.S.; Nuthmann, A.; Nyström, M.; Andersson, R.; Niehorster, D.C.; Hooge, I.T.C. The fundamentals of eye tracking part 1: The link between theory and research question. *Behav. Res. Methods* **2025**, *57*, 16. [CrossRef] [PubMed]
26. Jakobi, D.N.; Kern, T.; Reich, D.R.; Haller, P.; Jäger, L.A. PoTeC: A German naturalistic eye-tracking-while-reading corpus. *Behav. Res. Methods* **2025**, *57*, 211. [CrossRef] [PubMed]
27. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **1995**, *57*, 289–300. [CrossRef]
28. Hienert, D.; Schmidt, H.; Krämer, T.; Kern, D. EyeLiveMetrics: Real-time analysis of online reading with eye tracking. In *Proceedings of the 2024 Symposium on Eye Tracking Research and Applications (ETRA 2024), Glasgow, UK, 4–7 June 2024*; Association for Computing Machinery: New York, NY, USA, 2024; pp. 1–7. [CrossRef]
29. Mendhakar, A.; Sierak, K.; Płużyczka, M.; Leder, H. Interactive text reading: A text 2.0 implementation. *Kwart. Neofilol.* **2024**, *71*, 58–76. [CrossRef]
30. Serrano-Carot, M.; Angele, B.; Xu, H.; Vasilev, M.R. Webcams can be used to study eye movements during reading. *Behav. Res.* **2026**, *58*, 331. [CrossRef] [PubMed]
31. Meiri, Y.; Shubi, O.; Hadar, C.A.; Nitzav, A.K.; Berzak, Y. Déjà vu? Decoding repeated reading from eye movements. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, Vienna, Austria, 27 July–1 August 2025*; Association for Computational Linguistics: Vienna, Austria, 2025; pp. 19460–19482. [CrossRef]
32. Jakobi, D.N.; Reich, D.R.; Prasse, P.; Hofmann, J.M.; Bolliger, L.S.; Jäger, L.A. Eye-tracking-while-reading: A living survey of datasets with open library support. *Behav. Res. Methods* **2026**, *58*, 263. [CrossRef] [PubMed]
33. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.