Predicting L2 Eye Movements with Generalization and Reader-Specific Adaptation

Anonymous ACL submission

Abstract

We cast L2 reading as a predictive generalization problem: can word-by-word eye movements be predicted for unseen readers and unseen texts? Using a large naturalistic eye-tracking corpus, we model nine eye-movement measures as token-level targets and evaluate cross-domain generalization under three settings: unseen texts, unseen readers, and both. To test whether learner variables provide transferable information beyond identity shortcuts, we compare true learner profiles against a column-wise permutation control. We further study the sample efficiency of personalization via a two-stage k-shot residual calibration method that adapts a general predictor to new readers. Our framework establishes a reproducible benchmark for L2 reading prediction and links cognitive variables to out-of-sample performance.

1 Introduction

Predicting human eye movements from text has become an increasingly powerful route for linking cognitive theory with data-driven modeling. In first-language (L1) reading, eye-movement records—fixation durations, skipping, and regressions—provide dense token-level behavioral signals that correlate with lexical access, attentional allocation, and saccade planning (Reichle et al., 2003; Engbert et al., 2005). Classic models such as E-Z Reader explain variation in reading time via serial attention (Reichle et al., 2003), while dynamical models such as SWIFT emphasize parafoveal preview and parallel processing (Engbert et al., 2005; Seelig et al., 2021). At the same time, advances in corpus-based and information-theoretic methods have revealed strong effects of frequency, length, and predictability on reading time (Shain, 2024), and surprisal from neural language models now accounts for substantial variance across languages (Hao et al., 2020; Gotlieb Wilcox et al., 2023; Hofmann et al., 2022; Wiechmann et al., 2022; Bianchi et al., 2025). As a result, "predicting token-level eye-movement responses given text and context" has emerged as a computationally meaningful task with both cognitive and NLP relevance.

Research on second-language (L2) reading, however, presents additional challenges that are rarely addressed from a predictive modeling standpoint. Prior work documents consistent group-level differences: L2 readers exhibit longer fixations, more regressions, and heightened sensitivity to lexical and syntactic complexity (Gotlieb Wilcox et al., 2023; Hofmann et al., 2022; Shain, 2024; Wiechmann et al., 2022; Bianchi et al., 2025). Yet demonstrating an effect does not guarantee that it yields *generalizable* predictions. When readers or texts overlap between train and test, models can capture idiosyncratic statistics of individuals or corpora, inflating apparent performance while masking failures under genuine out-of-distribution conditions. A central, underexplored question is therefore: *How well can we predict L2 eye-movement behavior for unseen readers and unseen texts?*

A second challenge lies in heterogeneity among L2 learners. Variables such as proficiency, vocabulary size, and working memory correlate with unfamiliar-word processing, regression behavior, and saccade patterns (Quiñonez-Beltran et al., 2024; Berzak et al., 2022). However, existing evidence is largely correlational. From a predictive perspective, two key questions remain open: (i) Do learner variables offer *transferable* benefits across new readers and new texts, or do they mainly function as identity cues that help only when reader overlap leaks into evaluation? (ii) For a target reader, what is the relative value of learner variables versus a *small amount* of that reader's own eye-movement data?

Within NLP, work on reading-related prediction has framed tasks such as comprehension prediction (Dirix et al., 2020) and gaze–text modeling (Hao et al., 2020; De Varda and Marelli, 2022),

but benchmarks typically focus on L1 reading or adopt coarse evaluation setups. Crucially, there is no systematic assessment of multi-dimensional token-level eye-tracking measures in L2 reading under strict out-of-sample generalization. Meanwhile, few-shot personalization methods—e.g., residual adaptation and lightweight tuning (Whitford and Titone, 2016; Nahatame, 2023)—indicate that small samples may suffice for personalization, yet they have not been used to quantify a basic empirical question: *How many eye-movement samples does personalization need to meaningfully improve predictions for a new L2 reader?*

To address these gaps, we introduce a unified token-level prediction framework for L2 reading organized around three falsifiable research questions (RQs). Using a large-scale natural-reading corpus, we formulate nine eye-tracking measures as token-level prediction tasks and evaluate cross-domain generalization under three complementary splits: TEXT (unseen texts), READER (unseen readers), and RXT (unseen readers + unseen texts). We focus on:

- **RQ1 (cross-domain generalization):** How does token-level predictive performance change when holding out readers or texts, and when holding out both? Which eye-movement measures are most vulnerable to cross-domain failures?
- **RQ2 (transferability of learner variables):** How much predictive gain do learner variables provide under in-sample versus out-of-sample conditions? To what extent are these gains attributable to transferable information versus superficial identity shortcuts?
- **RQ3 (sample efficiency of personalization):** Under cross-reader evaluation, how much improvement can be obtained from k eye-movement samples from the target reader, and what does the performance–k curve look like?

To examine RQ2, we introduce a tightly controlled diagnostic: using identical models and splits, we compare three input settings—`w` (lexical/context features), `all` (adding learner variables), and `all-permute` (column-wise permutation of learner variables)—to disentangle genuine transfer from identity leakage. To examine RQ3, we propose a two-stage k-shot residual calibration procedure: a general predictor is trained first, followed by lightweight residual fitting on k samples from the target reader, enabling principled personalization without architectural modification.

Our contributions are threefold. (i) We systematize a token-level out-of-sample prediction setting for L2 reading, covering nine measures and three orthogonal generalization splits. (ii) We propose a diagnostic framework that quantifies the transferability of learner variables and exposes identity-related leakage risks. (iii) We provide the first empirical analysis of k-shot personalization for L2 reading, characterizing the marginal value of small individual samples and offering a reproducible evaluation benchmark for future work.

2 Related Work

2.1 Models of Eye Movements in Reading

Mechanistic models of eye-movement control for L1 reading are well established. E-Z Reader models largely serial lexical processing and attention shifts, capturing fixation durations and skipping (Reichle et al., 2003). SWIFT assumes parallel activation with foveal inhibition and stochastic saccade timing, supported across populations and writing systems (Engbert et al., 2005; Laubrock et al., 2006; Zang, 2019), with Bayesian parameter estimation for individual differences (Seelig et al., 2020; Rabe et al., 2019) and extensions modeling parafoveal preview effects (Seelig et al., 2021). However, these approaches are primarily L1-focused and simulation-driven, with limited out-of-sample prediction and evaluation across *multiple* eye-movement metrics under a unified evaluation protocol.

2.2 Linguistic Predictors of Reading Time

Information-theoretic predictors, especially neural surprisal, explain word-level reading times across languages and in L2, while perplexity is not a linear proxy for predictivity (Hao et al., 2020; De Varda and Marelli, 2022; Gotlieb Wilcox et al., 2023). Surprisal often outperforms cloze probability (Hofmann et al., 2022); frequency and predictability can be temporally dissociated (Shain, 2024); and Transformer models capture nonlinear gaze and regression patterns in naturalistic reading (Wiechmann et al., 2022; Bianchi et al., 2025). Yet evaluations typically remain L1-centric and single-corpus/metric, leaving generalization across texts (TEXT), readers (READER), and joint shifts (RXT) underexplored.

Table 1: Background statistics across four L2 groups (Mean(SD) / Range).

	L (N=74)		T (N=67)		V (N=67)		M (N=53)	
Variable	M(SD)	Range	M(SD)	Range	M(SD)	Range	M(SD)	Range
age	23.32 (4.18)	19–41	22.64 (4.97)	18–42	23.66 (4.26)	19–43	23.32 (3.91)	18–35
learner_HSK	4.39 (0.54)	3–6	4.21 (0.69)	3–6	4.79 (0.77)	3–6	4.43 (0.69)	4–6
education_time	16.49 (4.28)	1–25	17.52 (4.82)	2–30	18.31 (3.13)	12–30	16.15 (3.16)	4–25
learning_chinese_age	17.93 (5.26)	3–32	15.94 (5.45)	1–36	18.76 (3.29)	7–27	18.32 (6.82)	1–34
learn_chinese_time	4.57 (3.46)	1–17	5.67 (4.20)	1–20	4.16 (2.71)	1–16	3.66 (4.03)	1–20
oral_age	17.65 (5.63)	2–32	17.01 (5.01)	1–35	19.03 (3.34)	7–27	18.94 (6.75)	1–34
talk_well_age	20.32 (4.17)	2–38	18.96 (4.27)	3–36	21.09 (3.04)	15–29	20.92 (5.48)	5–35
learning_read_age	18.50 (5.06)	2–32	17.21 (4.62)	3–36	19.57 (2.85)	15–28	19.62 (5.75)	5–34
read_well_age	20.47 (4.52)	2–38	19.30 (4.33)	3–36	21.48 (3.00)	15–29	21.09 (5.47)	5–35
chinese_oral	6.23 (1.32)	3–9	6.15 (1.52)	2–10	6.13 (1.31)	3–10	5.96 (1.39)	3–9
understand_chinese_oral	6.73 (1.36)	3–9	6.43 (1.44)	2–9	6.66 (1.23)	4–10	6.77 (1.41)	4–10
read_chinese	6.36 (1.39)	2–9	5.99 (1.57)	2–9	6.52 (1.20)	4–10	6.51 (1.58)	4–10

2.3 L2 Reading: Individual Differences and Prediction

L2 reading shows longer fixations, more regressions, and reduced perceptual span, modulated by proficiency and language distance (Quiñonez-Beltran et al., 2024; Berzak et al., 2022; Dirix et al., 2020; Whitford and Titone, 2016). Textual and cognitive factors (e.g., complexity, dependency distance, vocabulary, working memory) further shape difficulty and eye-movement patterns (Nahatame, 2023; Huang et al., 2022; Gnetov and Kuperman, 2024; Demareva and Edeleva, 2020; Leon Guerrero et al., 2021; Pan and Lin, 2020; El-Dakhs et al., 2024). Predictive work using gaze features for comprehension suggests cross-reader transfer but often uses coarse features and lacks token-level, multi-metric modeling (Gregg and D'Mello, 2018; Southwell et al., 2023; Shubi et al., 2024). Parameter-efficient/few-shot adaptation improves cross-domain prediction in other settings (Song et al., 2023; Gao et al., 2025), but is rarely studied for L2 eye-movement prediction.

3 Method

3.1 Data Source and Participants

We use the Chinese Learner Eye-tracking Corpus (CECO) (Lu et al., 2025), which includes Chinese L2 readers from Laos (L, $N = 74$), Thailand (T, $N = 67$), Vietnam (V, $N = 67$), and Myanmar (M, $N = 53$), totaling $N = 261$ participants. Overall proficiency is intermediate to high, with group mean HSK levels ranging from approximately 4.21 to 4.79 and self-rated speaking, listening, and reading skills averaging around 6/10. The four groups are broadly comparable in age, education, and learning experience (Table 1). CECO provides participant metadata and eye-movement events aligned to interest areas (IAs), from which we derive lexical and eye-movement measures, apply secondary filtering, and fit statistical models.

Table 2: Nine eye–movement indicators grouped by processing stage. Each token is one interest area (IA).

Indicator	**Definition (token-level)**
Early-stage (initial lexical processing)	
FFD	first fixation duration in IA
FRD	sum of fixations in IA during first run
FRFC	fixation count in IA during first run
Late-stage (regression/integration)	
RIC	times IA is entered from a later IA
RPD	fixations until eyes progress to later IA
DWL	sum of fixations across all runs in IA
Global (overall reading behavior)	
FIX%	proportion of all fixations falling in IA
SKIP	IA receives zero first-pass fixation (1=yes)
RUN	number of times IA is entered/exited

Table 3: Split statistics for each dataset.

Dataset	Train (K)	Val (K)	Test (K)	Split (%)
TEXT	3834.8	1144.7	1145.8	62.61/18.69/18.71
READER	3851.5	1038.5	1235.3	62.88/16.95/20.17
RXT	3927.6	1068.4	1243.1	62.95/17.12/19.93

3.2 Reading Materials and Lexical Features

Reading materials. Materials are drawn from the CECO natural discourse corpus (Lu et al., 2025) and comprise 60 passages (20 narrative, 20 expository, 20 argumentative) covering diverse everyday and academic topics. Passage length ranges from 46 to 232 characters, with 70.8% of words being disyllabic. Overall lexical difficulty targets intermediate L2 readers: HSK1–4 account for 84.8% of tokens, and HSK5, HSK6, and out-of-level items

for 3.8%, 2.9%, and 8.5%, respectively.

Lexical features. Each word token is annotated with fourteen lexical, psycholinguistic, and orthographic features: LOG_FREQUENCY, LENGTH, SURPRISAL, FAMILIARITY, TRANSPARENCY, IMAGEABILITY, CONCRETENESS, VALENCE, AROUSAL, STROKES, RADICAL, MEANING, and HSK.

LOG_FREQUENCY is log-transformed frequency from the SUBTLEX-CH subtitle corpus (Cai and Brysbaert, 2010); LENGTH is the number of characters. SURPRISAL is trigram surprisal computed with an add-k-smoothed 3-gram language model trained split-wise on training passages to avoid text-level leakage (Appendix 9.1). HSK encodes lexical proficiency level under the *International Standard for Chinese Language Proficiency*; the construction of this mapping and related lexical norms are described in Appendix 9.1. According to CECO norms, 74.7% of words have familiarity ≥ 6 and 91.6% have familiarity ≥ 5.

Psycholinguistic ratings come from both L2 learners and native speakers: FAMILIARITY reflects L2 learners' subjective familiarity, whereas TRANSPARENCY, IMAGEABILITY, CONCRETENESS, VALENCE, and AROUSAL are based on native-speaker norms and additional rating studies. Orthographic and morphological properties are derived from lexical resources: STROKES is the total number of character strokes, RADICAL is the number of components/radicals, and MEANING is the number of dictionary senses in the *Modern Chinese Dictionary*. Full details of the lexical norms, lexical resources, and annotation pipeline, including dictionary sources and URLs, are provided in Appendix 9.1.

3.3 Eye-tracking Data Collection and Preprocessing

Eye-movement data are taken from the CECO eye-tracking database (Lu et al., 2025). Recordings were made with an EyeLink 1000 Plus (1500 Hz, tower-mount). Passages were presented in full on a 23.8-inch monitor in 22-pt Songti font; participants read silently at their own pace and then answered comprehension questions. A standard 9-point calibration with drift correction preceded each trial.

3.4 Tasks and Input Representation

We treat each segmented token as an interest area (IA) and predict nine IA-level eye-movement measures as separate targets y_k ($k = 1, \ldots, 9$). The targets span processing stages from early local/duration-based measures (e.g., FFD/FRD/FRFC, DWL, FIX%) to regression- and strategy-related behavior (e.g., RIC/RPD/RUN). Continuous and nonnegative/count-valued targets are modeled with regression objectives (including Poisson regression); IA_SKIP is treated as a 0/1-valued regression target, and all models are evaluated using R^2.

Each IA is represented as

$$\mathbf{x} = (\mathbf{x}_{\text{word}}, \mathbf{x}_{\text{reader}}),$$

where $\mathbf{x}_{\text{word}}$ contains 14 word-level lexical features and $\mathbf{x}_{\text{reader}}$ contains 13 reader background variables. All experiments below use the same targets and base representation; because background variables may act as reader-indexing cues when readers overlap between training and testing, we include a permutation control (Experiment 2) and test-time personalization (Experiment 3).

3.5 Experimental Design

Splits and model classes. We evaluate generalization with three splits: TEXT holds out texts (reader overlap), READER holds out readers (text overlap), and RXT holds out both texts and readers; all splits are applied within each of the four country groups (Table 3). Across experiments, we compare ElasticNet (Zou and Hastie, 2005), Poisson regression (Nelder and Wedderburn, 1972), LightGBM (Ke et al., 2017), and a Transformer encoder(Vaswani et al., 2017). Training, optimization, and permutation/resampling procedures are detailed in Appendix 9.2.

Experiment 1: Word-feature transfer. The first experiment isolates how far lexical information transfers across texts and readers. For each split $d \in \{\text{TEXT}, \text{READER}, \text{RXT}\}$ and model class m, we train on the *word* configuration and evaluate on the test set $\mathcal{T}_d$. We report

$$R^2_{m,d} = 1 - \frac{\sum_{i \in \mathcal{T}_d} (y_i - \hat{y}_i^{(m)})^2}{\sum_{i \in \mathcal{T}_d} (y_i - \bar{y}_d)^2}, \tag{318}$$

$$\bar{y}_d = \frac{1}{|\mathcal{T}_d|} \sum_{i \in \mathcal{T}_d} y_i. \tag{319}$$

To control for split difficulty, we summarize transfer relative to TEXT via $R^2_{m,d} - R^2_{m,\text{TEXT}}$, i.e., the additional cost when shifting to held-out readers and/or texts+readers.

Table 4: **Token-level prediction performance (absolute R^2) across three generalization splits.** For each indicator and model, we report test-set R^2 under *w* (lexical features only) and *all* (lexical + learner/background features), along with the incremental gains $\Delta_{\text{all}} = R^2_{\text{all}} - R^2_{\text{w}}$ and $\Delta_{\text{perm}} = R^2_{\text{perm}} - R^2_{\text{w}}$.

		READER				RXT				TEXT			
Indicator	**Model**	w	all	Δall	Δperm	w	all	Δall	Δperm	w	all	Δall	Δperm
Early-stage (initial lexical processing)													
FFD	ElasticNet	0.008†	-0.003†	-0.011†	-0.001†	0.005†	-0.004†	-0.009†	-0.039†	0.009†	0.014†	+0.005†	+0.006†
	Transformer	0.001	-0.025†	-0.027†	-0.016†	-0.022†	-0.044†	-0.022†	-0.042†	-0.014†	0.011†	+0.025†	+0.020†
	LightGBM	0.025†	0.004†	-0.021†	-0.004†	0.005†	-0.011†	-0.016†	-0.021†	0.009†	0.080†	+0.070†	+0.071†
FRD	ElasticNet	0.136†	0.129†	-0.007†	-0.012†	0.123†	0.121†	-0.002	-0.000	0.116†	0.130†	+0.014†	+0.005†
	Transformer	0.180†	0.153†	-0.027†	-0.031†	0.092†	0.087†	-0.004	-0.025†	0.100†	0.112†	+0.012†	+0.002
	LightGBM	0.176†	0.145†	-0.031†	-0.029†	0.118†	0.102†	-0.016†	-0.046†	0.114†	0.186†	+0.072†	+0.071†
FRFC	Poisson	0.183†	0.193†	+0.010†	-0.005†	0.172†	0.186†	+0.014†	-0.219†	0.161†	0.176†	+0.015†	+0.004†
	Transformer	0.248†	0.231†	-0.017†	-0.036†	0.161†	0.165†	+0.004	+0.004†	0.164†	0.176†	+0.012†	+0.001
	LightGBM	0.210†	0.199†	-0.011†	-0.047†	0.148†	0.154†	+0.006†	-0.011†	0.145†	0.199†	+0.054†	+0.055†
Late-stage (regression / integration processing)													
RIC	Poisson	0.030†	0.031†	+0.001†	-0.005†	0.032†	0.033†	+0.001	-0.004†	0.038†	0.040†	+0.002†	+0.003†
	Transformer	0.067†	0.030†	-0.037†	-0.024†	0.024†	0.013†	-0.011†	+0.005†	0.030†	0.031†	+0.002†	+0.008†
	LightGBM	0.052†	0.035†	-0.016†	-0.009†	0.017†	0.017	0.000	0.000	0.023†	0.059†	+0.036†	+0.034†
RPD	ElasticNet	0.020†	0.019†	-0.001†	-0.002†	0.020†	0.021†	+0.001	0.000	0.018†	0.022†	+0.004†	+0.002†
	Transformer	0.026†	0.022†	-0.004†	-0.011†	0.015†	0.009†	-0.006†	-0.010†	0.012†	0.018†	+0.006†	+0.000
	LightGBM	0.031†	0.019†	-0.012†	-0.015†	0.017†	0.007	-0.010†	0.000	0.015†	0.031†	+0.016†	+0.016†
DWL	ElasticNet	0.206†	0.203†	-0.003†	-0.001†	0.200†	0.204†	+0.004†	-0.005†	0.185†	0.207†	+0.022†	+0.004†
	Transformer	0.247†	0.193†	-0.054†	-0.026†	0.184†	0.094†	-0.089†	-0.052†	0.178†	0.233†	+0.054†	+0.054†
	LightGBM	0.244†	0.216†	-0.028†	-0.039†	0.196†	0.179†	-0.017†	-0.021†	0.187†	0.286†	+0.099†	+0.097†
Global (overall reading behavior)													
FIX%	ElasticNet	0.217†	0.217†	0.000	0.000	0.225†	0.225†	0.000	0.000	0.213†	0.213†	0.000	0.000
	Transformer	0.454†	0.446†	-0.008†	-0.007†	0.396†	0.395†	-0.001	-0.003†	0.365†	0.369†	+0.004†	+0.008†
	LightGBM	0.383†	0.374†	-0.009†	-0.009†	0.218†	0.216†	-0.002†	-0.004†	0.209†	0.211†	+0.002†	+0.001†
SKIP	ElasticNet	0.049†	0.047†	-0.003†	-0.063†	0.041†	0.039†	-0.001	-0.014†	0.051†	0.056†	+0.005†	+0.004†
	Transformer	0.092†	0.057†	-0.035†	-0.029†	0.037†	-0.060†	-0.097†	-0.051†	0.050†	0.055†	+0.005†	-0.003†
	LightGBM	0.071†	0.045†	-0.026†	-0.030†	0.042†	0.023†	-0.019†	-0.015†	0.052†	0.116†	+0.064†	+0.064†
RUN	Poisson	0.064†	0.062†	-0.002†	-0.003†	0.068†	0.070†	+0.002	-0.019†	0.071†	0.082†	+0.011†	+0.000
	Transformer	0.070†	0.043†	-0.027†	-0.016†	0.039†	0.027†	-0.012†	-0.016†	0.056†	0.070†	+0.014†	+0.025†
	LightGBM	0.055†	0.035†	-0.020†	-0.023†	0.031†	0.028†	-0.004	-0.013†	0.044†	0.109†	+0.065†	+0.064†

Experiment 2: Background-information transfer. The second experiment tests whether reader background features improve prediction and whether such gains transfer. For each model and split, we train three configurations: *word* (word-only), *all* (word + background), and *all-permute* (word + background with reader-level strict permutation). We quantify gains over *word* as

$$\Delta_{\text{all}} = R^2_{\text{all}} - R^2_{\text{word}}, \qquad \Delta_{\text{perm}} = R^2_{\text{perm}} - R^2_{\text{word}},$$

where *perm* denotes the permutation condition. Comparing Δ_{all} vs. Δ_{perm} separates transferable alignment from gains consistent with reader-indexing cues; identity-leakage analyses are in Appendix 9.3 and protocol details in Appendix 9.2.

Experiment 3: k-shot reader adaptation. The third experiment evaluates sample-efficient test-time personalization. For READER and RXT, we start from a word-only LightGBM model and apply two-stage residual k-shot calibration: using k randomly sampled test instances per reader, we estimate a reader-specific residual bias and apply a shrinkage correction to all predictions. We consider multiple k-shot ratios and average over repeated subsampling runs; the full procedure is in Appendix 9.2.

Evaluation and uncertainty. For all experiments, $R^2_{m,d}$ is computed on held-out test sets. We summarize transfer via $R^2_{m,d} - R^2_{m,\text{TEXT}}$ and estimate 95% CIs by percentile bootstrapping on the

test set. For feature gains we use paired bootstrapping over test instances, and for k-shot adaptation we first average predictions over repeats per instance before applying the same resampling; see Appendix 9.2 for details.

4 Results

4.1 Experiment 1: Cross-domain generalization (RQ1)

We first ask how well word-level information transfers across held-out texts and readers, and which eye-movement measures are most affected (Table 4). Using the *word*-only configuration as reference, we compute $R^2_{m,d}$ for each split $d \in \{\text{TEXT}, \text{READER}, \text{RXT}\}$ and summarize transfer via $R^2_{m,d} - R^2_{m,\text{TEXT}}$.

Across models and splits, we observe a robust dissociation by processing stage. Early local measures (FFD, FRD, FRFC) and duration-based measures (DWL, FIX%) retain moderate out-of-domain predictability, whereas regression-related measures (RIC, RPD) and global strategy-related behavior (RUN) show lower R^2 and higher split sensitivity. For example, under READER, the Transformer achieves $R^2 \approx 0.18$–0.25 on FRD/FRFC but near-zero scores on RPD, and under RXT even FFD can become slightly negative. FIX% remains comparatively stable, while RUN is both weaker and more shift-sensitive.

Takeaway (RQ1). With lexical features only, local and duration-based measures transfer relatively well across domains, whereas regression-path and strategy-related behavior degrade more strongly, suggesting that word-level information primarily supports generalizable local and temporal processes.

4.2 Experiment 2: Transferability of reader background variables (RQ2)

Experiment 2 tests whether reader background features provide transferable gains and whether they act as reader-specific indices. We compare *all* (word + background) with *all-permute* (background strictly permuted across readers), focusing on Δ_{all} and Δ_{perm} (Table 4) and LightGBM radar plots (Figure 1).

Under READER and RXT, background features do not yield reliable improvements and often induce negative transfer; moreover, Δ_{perm} typically mirrors Δ_{all} in sign and magnitude. For instance, DWL degrades under both splits with similar drops

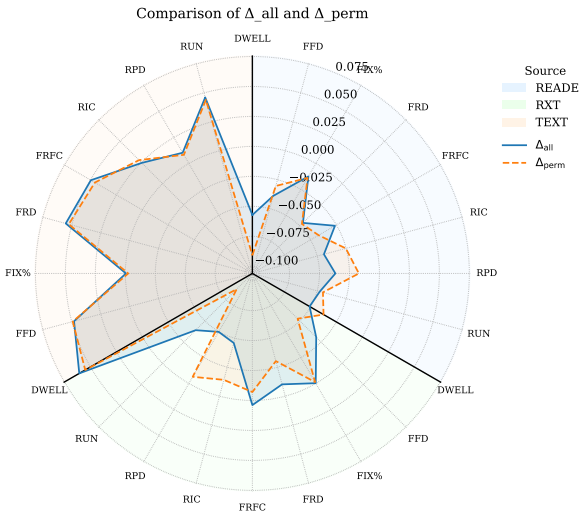

Figure 1: **LightGBM (main) radar comparison of Δ_{all} vs. Δ_{perm} across splits.** Each axis corresponds to an eye-movement indicator. Under READER and RXT, both curves are contracted and largely overlapping, indicating that learner/background features provide no transferable gains beyond lexical features. Under TEXT, curves remain overlapping but shift outward, consistent with identity-indexing effects rather than generalizable learner structure.

for *all* and *all-permute*, and the corresponding radar curves are contracted and nearly overlapping. In contrast, under TEXT (reader overlap), background features produce consistent gains, but these gains are largely unchanged by permutation: for LightGBM, DWL and RUN improve by almost identical amounts under *all* and *all-permute*, and the radar curves shift outward while remaining aligned.

This split-dependent pattern is consistent with background variables acting as non-transferable reader-specific cues rather than transferable individual-difference signals. Concretely, Appendix 9.3 shows that (i) background vectors exhibit statistically reliable coupling with one-hot reader identities under a permutation test (linear CKA, p=0.002), yet (ii) reader identity is highly recoverable from background variables alone using a simple logistic-regression probe (88.9% accuracy; random baseline $\approx 0.34\%$). Together with the permutation control, these results suggest that gains under TEXT are dominated by reader-specific indexing cues that are available only when readers overlap between training and testing.

Takeaway (RQ2). Reader background variables do not transfer under held-out readers/texts and mainly help when readers overlap, where gains are indistinguishable from the permutation con-

Table 5: LightGBM R^2 and ΔR^2 (vs. 0%) for k-shot calibration under READER and RXT splits. The 0% column corresponds to the uncalibrated *word-only* LightGBM model. For each calibration budget (0.5%–10%), we report absolute R^2 and the gain ΔR^2 relative to the 0% baseline (in parentheses). Colors indicate positive (green) or negative (red) changes. † marks statistically significant differences (95% CI excluding zero; CI omitted for brevity).

Ind.	0%	0.5%	1%	2%	5%	10%
READER split						
FFD	+0.025†	+0.068† (+0.043†)	+0.081† (+0.056†)	+0.088† (+0.063†)	+0.093† (+0.069†)	+0.095† (+0.070†)
FRD	+0.176†	+0.208† (+0.032†)	+0.217† (+0.041†)	+0.224† (+0.048†)	+0.227† (+0.051†)	+0.228† (+0.052†)
FRFC	+0.232†	+0.253† (+0.021†)	+0.260† (+0.027†)	+0.264† (+0.031†)	+0.267† (+0.035†)	+0.268† (+0.035†)
RIC	+0.072†	+0.087† (+0.015†)	+0.092† (+0.020†)	+0.096† (+0.024†)	+0.097† (+0.025†)	+0.098† (+0.026†)
RPD	+0.031†	+0.036† (+0.005†)	+0.039† (+0.008†)	+0.040† (+0.009†)	+0.042† (+0.011†)	+0.042† (+0.011†)
DWL	+0.244†	+0.281† (+0.037†)	+0.288† (+0.044†)	+0.294† (+0.050†)	+0.298† (+0.053†)	+0.298† (+0.054†)
FIX%	+0.383†	+0.382† (-0.000†)	+0.382† (-0.000)	+0.383† (+0.000†)	+0.383† (+0.001†)	+0.384† (+0.001†)
SKIP	+0.071†	+0.103† (+0.032†)	+0.115† (+0.044†)	+0.121† (+0.050†)	+0.125† (+0.054†)	+0.126† (+0.055†)
RUN	+0.090†	+0.114† (+0.025†)	+0.118† (+0.029†)	+0.124† (+0.034†)	+0.126† (+0.037†)	+0.127† (+0.037†)
RXT split						
FFD	+0.005†	+0.020† (+0.015†)	+0.034† (+0.029†)	+0.044† (+0.039†)	+0.068† (+0.063†)	+0.073† (+0.068†)
FRD	+0.118†	+0.129† (+0.011†)	+0.141† (+0.023†)	+0.150† (+0.032†)	+0.164† (+0.046†)	+0.172† (+0.054†)
FRFC	+0.171†	+0.178† (+0.008†)	+0.185† (+0.014†)	+0.193† (+0.022†)	+0.204† (+0.033†)	+0.210† (+0.039†)
RIC	+0.042†	+0.046† (+0.004†)	+0.050† (+0.008†)	+0.055† (+0.014†)	+0.065† (+0.023†)	+0.067† (+0.026†)
RPD	+0.017†	+0.017† (+0.001)	+0.022† (+0.005†)	+0.024† (+0.007†)	+0.028† (+0.012†)	+0.030† (+0.013†)
DWL	+0.196†	+0.208† (+0.012†)	+0.218† (+0.021†)	+0.229† (+0.033†)	+0.241† (+0.045†)	+0.250† (+0.054†)
FIX%	+0.220†	+0.221† (+0.001†)	+0.222† (+0.002†)	+0.223† (+0.003†)	+0.226† (+0.006†)	+0.227† (+0.007†)
SKIP	+0.042†	+0.056† (+0.014†)	+0.069† (+0.027†)	+0.085† (+0.043†)	+0.102† (+0.060†)	+0.109† (+0.067†)
RUN	+0.073†	+0.084† (+0.012†)	+0.086† (+0.014†)	+0.097† (+0.024†)	+0.108† (+0.035†)	+0.112† (+0.039†)

trol. Combined with strong identity recoverability from background variables (Appendix 9.3), this pattern supports a reader-indexing interpretation rather than transferable individual-difference structure, cautioning against naive use of raw background features in cross-domain settings.

4.3 Experiment 3: k-shot reader adaptation (RQ3)

Experiment 3 evaluates the sample efficiency of test-time k-shot personalization. Starting from a word-only LightGBM model, we apply residual calibration on READER and RXT. Table 5 reports absolute R^2 and ΔR^2 relative to k=0%; Figure 2 shows budget–response curves. Absolute R^2 values with 95% CIs for non-zero budgets, reported in Appendix 9.4, mirror the same budget–response pattern observed for ΔR^2.

Under READER, most measures exhibit steep early gains followed by tapering. For FFD, R^2 increases from $\approx$ 0.03 at 0% to $\approx$ 0.08 at 1% and $\approx$ 0.10 at 10%, with similar "early jump then plateau" trends for FRD and DWL. FIX% is largely insensitive to calibration (near-zero ΔR^2) but already has high baseline R^2.

Under RXT, calibration remains beneficial but more gradual: FFD, DWL, and SKIP all show steady improvements up to 10%, and even FIX% gains slightly, though absolute magnitudes remain modest.

Takeaway (RQ3). Across both splits, k-shot personalization yields strong gains at very low budgets (0.5–1%), slower improvements around 2%, and diminishing returns beyond 5%. In practice, small calibration budgets already recover a substantial fraction of the attainable personalization benefit.

5 Discussion

Experiments 1–3 clarify what transfers in cross-domain L2 eye-movement prediction and where models break. Lexical features generalize best for early-stage and duration-based measures, whereas regression-path and strategy-related indicators drop sharply under held-out readers and especially under joint reader–text shift. This dissociation suggests that word-level predictors capture relatively stable local processing regularities, while regressions and global strategies depend more on reader- and context-specific control.

Learner/background variables offer little out-of-domain benefit. Under reader overlap (TEXT), they appear helpful, but the near-identical gains of *all* and the strict *all-permute* control indicate that improvements largely reflect reader indexing

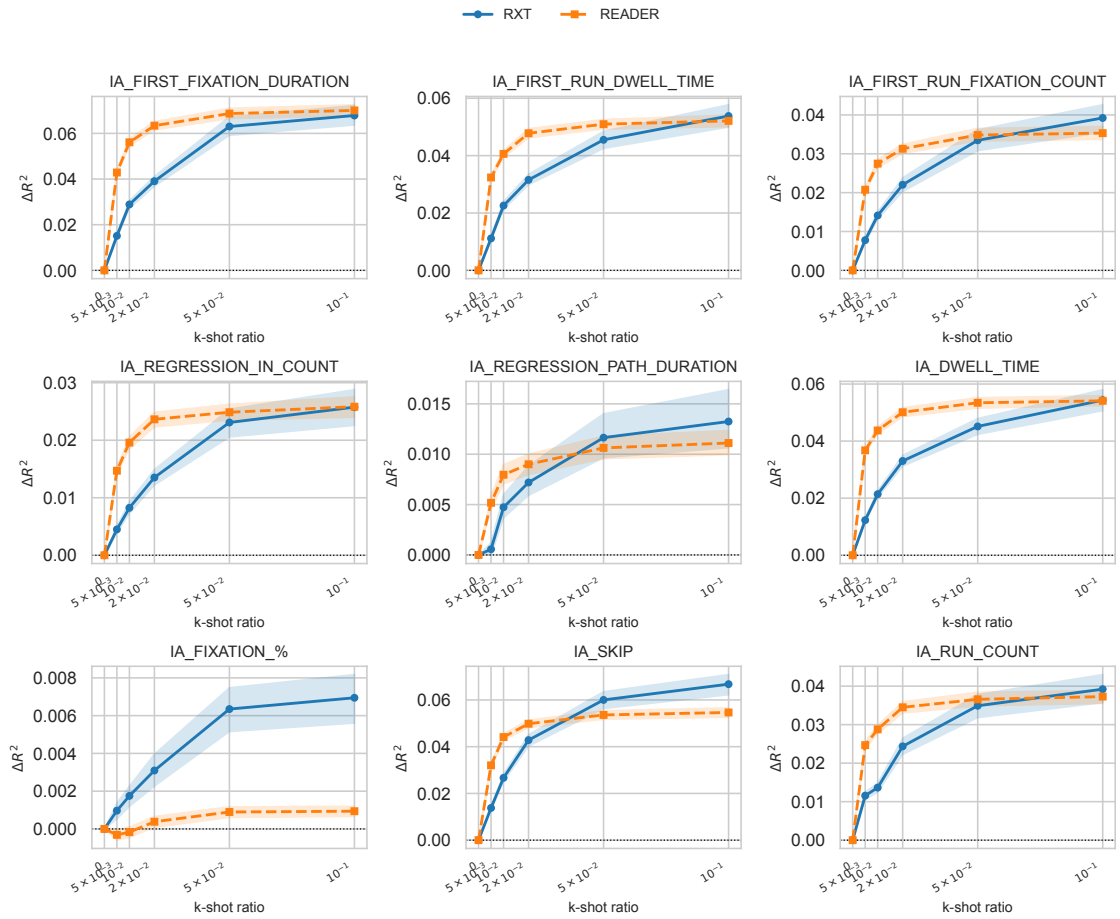

Figure 2: K-shot calibration of LightGBM with centered R^2 (95% CI) on READER and RXT. Most gains occur within 0.5%–1% reader data, then saturate. Error bars show bootstrap 95% CIs.

rather than transferable learner effects. Identity analyses support this interpretation: modest CKA can coexist with high identity decodability because CKA measures global geometric alignment, whereas a simple probe can exploit local separability in background-feature space. Practically, this cautions that raw profiles can inflate performance whenever reader overlap leaks into evaluation.

By contrast, lightweight residual k-shot calibration is a sample-efficient way to capture non-transferable variance. Very small budgets yield large early gains with diminishing returns, and gains are strongest for measures that are otherwise hard to transfer. Methodologically, these findings motivate strict cross-reader evaluation and explicit anti-leakage diagnostics (e.g., permutation controls, identity probes). Substantively, they favor hybrid approaches: strong lexical predictors for population-level regularities combined with de-identified personalization or hierarchical individual-effect modeling, alongside extensions that add richer linguistic structure while preserving rigorous domain separation.

6 Conclusion

We cast L2 reading as token-level prediction and evaluate nine eye-tracking measures under TEXT, READER, and RXT. With lexical/context features (w), duration/local measures generalize relatively well, while regression/strategy measures degrade under domain shift. Learner variables help only with reader overlap (TEXT) but fail under READER/RXT and may hurt; permutation yields $\Delta_{\text{all}} \approx \Delta_{\text{perm}}$, indicating identity cues rather than transferable effects. We therefore recommend de-identified ability/strategy representations or hierarchical (partial-pooling) individual-effect modeling for cross-domain prediction. Finally, k-shot residual calibration is sample-efficient: the largest gains occur at $k = 0.5\%$–1%, growth slows near $k \approx 2\%$, and returns diminish by $k \geq 5\%$ (with more headroom in RXT); FIX% is not harmed and is largely saturated in-domain.

7 Limitations

Using naturalistic reading data from learners with four Southeast Asian L1 backgrounds, we conducted a systematic analysis of cross-reader and cross-text eye-movement prediction. Although the participant population is relatively concentrated, we argue that the shared learning experiences and proficiency distributions associated with these language backgrounds help reduce confounds introduced by cross-group heterogeneity. Nevertheless, broader validation across more diverse L1 populations is needed. Moreover, our modeling efforts focus on lexical features and learner variables; other layers of linguistic information (e.g., discourse- or syntax-level structure) represent natural extensions for future work rather than core objectives of the current study. Finally, while the adopted k-shot calibration procedure performs robustly under laboratory conditions, its behavior in richer application scenarios warrants further investigation.

8 Ethics Statement

All eye-movement data used in this study were collected from voluntary participants with full informed consent under the original project protocols, and all data were anonymized for the present study. Demographic and background variables were deidentified and contain no personally identifiable information. The study does not involve automated decision-making or high-risk applications; the predictive models are intended solely to advance understanding of L2 reading processes. We make no value judgments about individuals or groups, and all statistical analyses aim to model behavioral variation rather than evaluate learning ability. Future work will continue to follow data-privacy and research-ethics guidelines, maintaining transparency and caution as data collection and model applications expand.

References

Yevgeni Berzak, Chie Nakamura, Amelia Smith, Emily Weng, Boris Katz, Suzanne Flynn, and Roger Levy. 2022. Celer: A 365-participant corpus of eye movements in l1 and l2 english reading. *Open Mind*, 6:41–50.

Bruno Bianchi, Fermín Travi, and Juan E Kamienkowski. 2025. Modeling cognitive processes of natural reading with transformer-based language models. *arXiv preprint arXiv:2505.11485*.

Qing Cai and Marc Brysbaert. 2010. Subtlex-ch: Chinese word and character frequencies based on film subtitles. *PloS one*, 5(6):e10729.

Andrea De Varda and Marco Marelli. 2022. The effects of surprisal across languages: Results from native and non-native reading. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 138–144.

Valeriia Demareva and Yulia Edeleva. 2020. Eye-tracking based l2 detection: Universal and specific eye movement patterns in l1 and l2 reading. *Procedia Computer Science*, 169:673–676.

Nicolas Dirix, Heleen Vander Beken, Ellen De Bruyne, Marc Brysbaert, and Wouter Duyck. 2020. Reading text when studying in a second language: An eye-tracking study. *Reading Research Quarterly*, 55(3):371–397.

Dina Abdel Salam El-Dakhs, Suhad Sonbul, and Ahmed Masrai. 2024. An eye-tracking study on the processing of l2 collocations: The effect of congruency, proficiency, and transparency. *Journal of Psycholinguistic Research*, 53(2):30.

Ralf Engbert, Antje Nuthmann, Eike M Richter, and Reinhold Kliegl. 2005. Swift: a dynamical model of saccade generation during reading. *Psychological review*, 112(4):777.

Boyan Gao, Xin Wang, Yibo Yang, and David Clifton. 2025. Optimization-inspired few-shot adaptation for large language models. *arXiv preprint arXiv:2505.19107*.

Daniil Gnetov and Victor Kuperman. 2024. Reading proficiency predicts spatial eye-movement control in the first and second language. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 50(8):1315.

Ethan Gotlieb Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *arXiv e-prints*, pages arXiv–2307.

Julie Gregg and Sidney D'Mello. 2018. Predicting reading comprehension from eye gaze. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 40.

Yiding Hao, Simon Mendelsohn, Rachel Sterneck, Randi Martinez, and Robert Frank. 2020. Probabilistic predictions of people perusing: Evaluating metrics of language model performance for psycholinguistic modeling. *arXiv preprint arXiv:2009.03954*.

Markus J Hofmann, Steffen Remus, Chris Biemann, Ralph Radach, and Lars Kuchinke. 2022. Language models explain word reading times better than empirical predictability. *Frontiers in Artificial Intelligence*, 4:730570.

Lingshan Huang, Jinghui Ouyang, and Jingyang Jiang. 2022. The relationship of word processing with l2 reading comprehension and working memory: Insights from eye-tracking. *Learning and Individual Differences*, 95:102143.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

Jochen Laubrock, Reinhold Kliegl, and Ralf Engbert. 2006. Swift explorations of age differences in eye movements during reading. *Neuroscience & Biobehavioral Reviews*, 30(6):872–884.

Sibylla Leon Guerrero, Veronica Whitford, Laura Mesite, and Gigi Luk. 2021. Text complexity modulates cross-linguistic sentence integration in l2 reading. *Frontiers in Communication*, 6:651769.

Shiyi Lu, Yaolin Lao, Xiuting Zhang, Wenhua You, and Rui Sun. 2025. Building an eye-tracking corpus for chinese learners in the asean region: A cross-linguistic comparison of vietnamese, thai, and laotian native speakers. *Applied Linguistics*, pages 76–88.

Shingo Nahatame. 2023. Predicting processing effort during l1 and l2 reading: The relationship between text linguistic features and eye movements. *Bilingualism: Language and cognition*, 26(4):724–737.

John Ashworth Nelder and Robert WM Wedderburn. 1972. Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 135(3):370–384.

Jue Pan and Dan Lin. 2020. Visuospatial memory uniquely predicts chinese reading comprehension in hong kong typically developing kindergarteners. *Reading and Writing*, 33(9):2205–2221.

Juan Francisco Quiñonez-Beltran, Taylor Marissa Seymour, Robert AJ Robbins, Ying Xu, and R Malatesha Joshi. 2024. What can eye movements tell us about reading in a second language: A scoping review of the literature. *Education Sciences*, 14(4):375.

Maximilian M Rabe, Johan Chandra, André Krügel, Stefan A Seelig, and Ralf Engbert. 2019. Bayesian inference of the swift model: Reading mirrored, scrambled, and normal texts. *Journal of Eye Movement Research*, 12(7).

Erik D Reichle, Keith Rayner, and Alexander Pollatsek. 2003. The ez reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and brain sciences*, 26(4):445–476.

Stefan Seelig, Sarah Risse, and Ralf Engbert. 2021. Predictive modeling of parafoveal information processing during reading. *Scientific reports*, 11(1):12954.

Stefan A Seelig, Maximilian M Rabe, Noa Malem-Shinitski, Sarah Risse, Sebastian Reich, and Ralf Engbert. 2020. Bayesian parameter estimation for the swift model of eye-movement control during reading. *Journal of Mathematical Psychology*, 95:102313.

Cory Shain. 2024. Word frequency and predictability dissociate in naturalistic reading. *Open Mind*, 8:177–201.

Omer Shubi, Yoav Meiri, Cfir Avraham Hadar, and Yevgeni Berzak. 2024. Fine-grained prediction of reading comprehension from eye movements. *arXiv preprint arXiv:2410.04484*.

Yisheng Song, Ting Wang, Puyu Cai, Subrota K Mondal, and Jyoti Prakash Sahoo. 2023. A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. *ACM Computing Surveys*, 55(13s):1–40.

Rosy Southwell, Caitlin Mills, Megan Caruso, and Sidney K D'Mello. 2023. Gaze-based predictive models of deep reading comprehension. *User Modeling and User-Adapted Interaction*, 33(3).

Yongqiang Su, Yixun Li, and Hong Li. 2023. Imageability ratings for 10,426 chinese two-character words and their contribution to lexical processing. *Current Psychology*, 42(27):23265–23276.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Veronica Whitford and Debra Titone. 2016. Eye movements and the perceptual span during first-and second-language sentence reading in bilingual older adults. *Psychology and Aging*, 31(1):58.

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

Xu Xu, Jiayin Li, and Huilin Chen. 2022. Valence and arousal ratings for 11,310 simplified chinese words. *Behavior research methods*, 54(1):26–41.

Chuanli Zang. 2019. New perspectives on serialism and parallelism in oculomotor control during reading: The multi-constituent unit hypothesis. *Vision*, 3(4):50.

Hui Zou and Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320.

9 Appendix

9.1 Lexical norms and annotation details

This appendix provides additional details about the lexical norms and annotation procedures described in Section 3.2.

Corpus-based features

We annotated each word token in the CECO passages with several corpus-based features. LOG_FREQUENCY denotes the log-transformed word frequency of the corresponding type as reported in SUBTLEX-CH, a large-scale frequency database based on Chinese film subtitles (Cai and Brysbaert, 2010). LENGTH is word length measured in the number of Chinese characters.

To capture contextual predictability, we computed SURPRISAL using an add-k smoothed trigram language model with split-wise training to avoid text-level leakage under cross-text evaluation. Concretely, for each evaluation split $d \in \{\text{TEXT}, \text{READER}, \text{RXT}\}$, we train a separate trigram model LM_d *only* on the training-set passages of split d, and then compute surprisal for tokens in the corresponding validation and test partitions under LM_d. For each token w_i, surprisal is defined as $-\log p_{\text{LM}_d}(w_i \mid w_{i-2}, w_{i-1})$, where conditional probabilities are estimated from trigram counts with add-k smoothing. We use sentence-boundary markers and the natural logarithm. This procedure ensures that, when texts are held out (TEXT and RXT), no validation/test passages contribute to the trigram statistics underlying surprisal features.

HSK encodes the lexical proficiency level according to the *International Standard for Chinese Language Proficiency*.[1] Each word was mapped to its corresponding band in the official standard where possible.

L2 familiarity ratings

FAMILIARITY reflects subjective familiarity judgments from Chinese L2 learners. We recruited 55 L2 learners of Chinese to rate the words contained in the CECO corpus. The participants were drawn from multiple L1 backgrounds, including Vietnam, Thailand, Laos, Indonesia, Cambodia, and Myanmar. The sample comprised 11 males (20.0%) and 44 females (80.0%), with ages ranging from 18 to 39 years ($M_{\text{age}} = 23.5$, $SD = 3.69$). All participants were intermediate-to-advanced learners whose Chinese proficiency was at or above HSK Level 4, and their Chinese learning experience ranged from 1 to 23 years ($M = 8.05$, $SD = 5.93$).

Familiarity was rated on a 7-point Likert scale (1 = completely unfamiliar, 7 = very familiar), following standard procedures in lexical familiarity research. Each participant was assigned a subset of the target words, and the resulting ratings were averaged to yield a familiarity score per word type.

Native-speaker psycholinguistic norms

The remaining psycholinguistic features were based on ratings from native speakers of Mandarin Chinese and existing norming studies. IMAGEABILITY scores primarily came from Su et al.'s (2023) imageability norms for 10,426 two-character Chinese words (Su et al., 2023). VALENCE and AROUSAL ratings were taken from Xu et al.'s (2022) affective norms for 11,310 simplified Chinese words (Xu et al., 2022). CONCRETENESS and TRANSPARENCY were obtained from native-speaker ratings using 7-point Likert scales, following prior work on concreteness, semantic representation, and morphological transparency.

For lexical items in the CECO corpus that were not covered by these published resources (for IMAGEABILITY, CONCRETENESS, VALENCE, or AROUSAL), we collected additional ratings from native speakers. We recruited 409 native Mandarin speakers (124 males, 30.3%; 285 females, 69.7%; $M_{\text{age}} = 24.8$, $SD = 0.45$), the majority of whom had undergraduate or graduate-level education (96.8%). Each participant was assigned at least one non-overlapping word list; each list was rated by approximately 100 participants. All ratings used 7-point Likert scales consistent with the existing norming studies. After standard data screening (e.g., removal of inconsistent responders), the number of valid ratings per word ranged from 67 to 100.

Orthographic and morphological features

The remaining orthographic and morphological features were derived automatically from lexical resources. STROKES denotes the total number of character strokes across all characters in a word, computed from standard stroke-count dictionaries. RADICAL represents the number of components (radicals or sub-character units) in the word.

[1] http://www.moe.gov.cn/jyb_xwfb/gzdt_gzdt/s5987/202103/W020210329527301787356.pdf

MEANING corresponds to the number of dictionary senses associated with a word. We used the *Modern Chinese Dictionary*[2] as the reference and counted the number of distinct sense entries for each headword. This yields a coarse measure of polysemy at the lexical type level.

9.2 Models and Training Details

Feature configurations and strict permutation. For each split (READER, RXT, TEXT), we consider three input configurations: (i) *word*: word-level features only; (ii) *all*: word features plus reader background features; (iii) *all-permute*: word features plus strictly permuted reader background features.

To construct the *all-permute* condition, we first concatenate the train/validation/test partitions and identify each reader by a unique recording/session identifier. We then extract the reader-specific background vector, apply a single global random permutation over readers, and assign the permuted background profile back to all samples from the same reader, while keeping sample ordering and split boundaries fixed. This procedure preserves all structure except the alignment between a reader and their background profile. For IA_SKIP, we follow the same protocol, treating the 0/1 outcome as a regression target and computing R^2 as defined below.

Transformer baseline. The Transformer operates on word / interest-area tokens (one record per interest area). We do not apply character- or subword-level tokenization and do not use pretrained representations. Sequences are grouped by recording/session identifier and trial-condition label, and tokens are ordered by interest-area index. We use dynamic padding within each batch and a standard key-padding mask for PAD tokens. The model attends over the full sequence within each instance, with sinusoidal positional encodings and maximum length `max_len = 5000`.

We use a Pre-LN `TransformerEncoder` with `num_layers = 3`, `d_model = 128`, `nhead = 4`, `dim_feedforward = 256`, and `dropout = 0.15`, with GELU activations. Word features are linearly projected to `d_model`, summed with positional encodings, and passed through LayerNorm. When reader background features are used, we inject them via FiLM conditioning: the background vector generates (γ, β) parameters (bounded by `tanh`) that scale and shift token representations. To reduce over-reliance on background information, we apply background dropout with rate `bg_dropout = 0.10` during training only.

Training uses AdamW (`lr = 1e-4`, `weight_decay = 1e-2`) with `SmoothL1Loss(beta = 1.0)`, gradient clipping at `max_grad_norm = 1.0`, and AMP on GPU. The learning-rate schedule uses linear warmup for the first `warmup_ratio = 0.05` fraction of steps followed by cosine decay. Each model is trained for up to `epochs = 1000` with early stopping on validation loss (`patience = 30`) and rollback to the best checkpoint. Hyperparameters are fixed across all settings; model selection uses only the validation set, with random seed 42.

Classical regression and LightGBM baselines. For continuous targets, we use ElasticNet (`alpha = 0.001`, `l1_ratio = 0.5`, `max_iter = 20000`, `random_state = 42`); for count targets, we use PoissonRegressor (`alpha = 0.001`, `max_iter = 20000`). Both models are implemented as a standardization-plus-regressor pipeline, fitted on the training set and evaluated directly on the test set.

LightGBM serves as a nonlinear baseline. Input features are standardized. For continuous targets, we apply z-score normalization (using the training mean and variance, applied consistently to training/validation); for count targets, we apply `log1p`. Predictions are inverse-transformed back to the original target scale. For IA_SKIP, we train under a regression objective on the original 0/1 target without additional transformation.

LightGBM hyperparameters are fixed as: `learning_rate = 0.01`, `num_leaves = 50`, `feature_fraction = 0.8`, `bagging_fraction = 0.8`, `bagging_freq = 5`, `lambda_l1 = 0.5`, `lambda_l2 = 0.5`, `min_data_in_leaf = 50`. We train for up to `num_boost_round = 20000` with early stopping on the validation set (`early_stopping_rounds = 100`, monitoring RMSE) and retain the best-iteration model.

Two-stage residual k-shot adaptation. To assess reader-level test-time adaptation with limited observations per reader (k-shot), we perform a two-stage residual calibration on RXT and READER.

[2] https://bjzs.vdict.com.cn/SY

Stage 1 (base predictor). We train a word-only LightGBM regressor on the training split (with the same hyperparameters as above) and select the best number of boosting iterations via early stopping on the validation split. This model yields predictions $\hat{y}^{\text{word}}$ on the test set, from which we define per-sample residuals:

$$r_i = y_i - \hat{y}_i^{\text{word}}.$$

Test samples are grouped by reader using the `DATA_FILE` identifier, which is used only for grouping and k-shot sampling.

Stage 2 (shrinkage calibration). For each reader with test set $\mathcal{D}$ of size n, given a k-shot ratio k_{ratio}, we set:

$$k = \max\big(1, \lfloor n \cdot k_{\text{ratio}} \rfloor\big),$$

and uniformly sample without replacement k instances as the k-shot subset $\mathcal{K}$. We estimate the reader-specific residual bias as the mean residual:

$$\bar{r} = \frac{1}{k} \sum_{j \in \mathcal{K}} r_j,$$

and apply shrinkage

$$\lambda = \frac{k}{k + \tau_{\text{shrink}}}, \qquad \tau_{\text{shrink}} = 20.0, \qquad \tilde{r} = \lambda \bar{r}.$$

The calibrated prediction for each test sample of that reader is

$$\hat{y}_i^{k\text{-shot}} = \hat{y}_i^{\text{word}} + \tilde{r}, \qquad i \in \mathcal{D}.$$

When a reader has fewer than k samples (i.e., $n < k$), we set $k = n$ and compute $\bar{r}$ over all available samples. We consider $k_{\text{ratio}} \in \{0.005, 0.01, 0.02, 0.05, 0.1\}$ and, for each ratio, run `n_repeats = 10` independent repeats with random k-shot subsets, using random seed 42 and a deterministic RNG per domain.

Evaluation and confidence intervals. For each domain $d \in \{\text{READER}, \text{RXT}, \text{TEXT}\}$, target y, model m, and feature mode, we evaluate on the test set $\mathcal{T}_d$ and report the within-domain coefficient of determination

$$R^2_{m,d} = 1 - \frac{\sum_{i \in \mathcal{T}_d} \big(y_i - \hat{y}_i^{(m)}\big)^2}{\sum_{i \in \mathcal{T}_d} \big(y_i - \bar{y}_d\big)^2},$$

$$\bar{y}_d = \frac{1}{|\mathcal{T}_d|} \sum_{i \in \mathcal{T}_d} y_i.$$

We estimate 95% confidence intervals (CIs) for R^2 using percentile bootstrapping with `n_bootstrap = 1000`: we resample test indices with replacement, recompute R^2 for each bootstrap sample, and report the 2.5th and 97.5th percentiles.

To quantify the effect of background features, we report within-domain differences with respect to the word-only baseline:

$$\Delta_{\text{all}} = R^2_{\text{all}} - R^2_{\text{word}},$$
$$\Delta_{\text{perm}} = R^2_{\text{perm}} - R^2_{\text{word}}.$$

Here, *all* denotes word+background features and *perm* denotes the strict permutation condition. CIs for Δ are obtained via paired bootstrap: for each bootstrap resample b (with shared indices for both modes), we compute

$$\Delta(b) = R^2_{\text{other}}(b) - R^2_{\text{word}}(b),$$

and take the 2.5th and 97.5th percentiles over $\{\Delta(b)\}$.

For repeated subsampling settings (e.g., k-shot), we first average predictions over repeats for each test instance and then apply the same bootstrapping procedures to obtain CIs for R^2 and Δ.

9.3 Identity-Related Signal Analysis

Method

We investigate whether learner background variables encode *reader-specific* information that can be exploited as a shortcut under splits where reader identities overlap between training and testing, and whether such information is *transferable* across readers. We conduct two complementary reader-level analyses: (i) **structural similarity** between background variables and identity representations, and (ii) **identity recoverability** via a lightweight decoding probe.

Representations. For each reader r, we form a background-variable vector $\mathbf{z}_r \in \mathbb{R}^d$ (d=12), comprising age, HSK level, learning duration, and self-assessed listening/speaking/reading proficiency, among others. Stacking these vectors yields $\mathbf{Z} \in \mathbb{R}^{N \times d}$. We also form an identity representation matrix $\mathbf{U} \in \mathbb{R}^{N \times N}$ via one-hot encodings of reader IDs.

Structural coupling (linear CKA). We compute *linear centered kernel alignment* (CKA) between $\mathbf{Z}$ and $\mathbf{U}$ to quantify whether the two

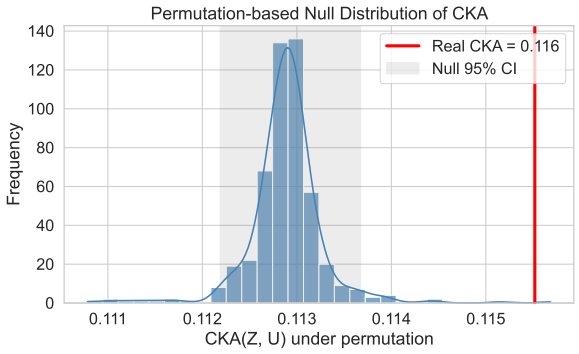

Figure 3: Permutation-based null distribution of linear CKA between reader background variables $\mathbf{Z}$ and one-hot identity representations $\mathbf{U}$. The red vertical line indicates the observed CKA on real data.

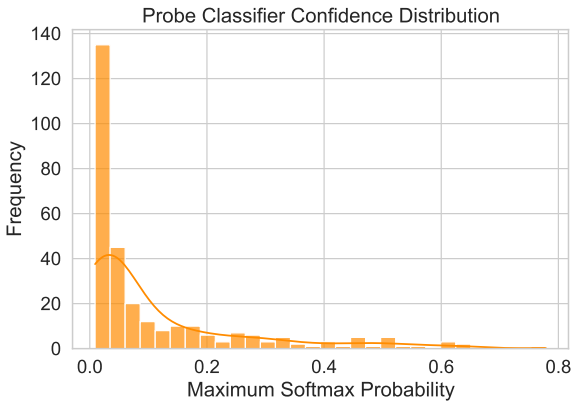

Figure 4: Distribution of probe classifier confidence (maximum softmax probability) when predicting reader identity from background variables.

spaces induce similar pairwise similarity structures across readers. Because CKA measures *global* geometric alignment (up to orthogonal transformations and isotropic scaling), a high CKA would suggest that background variables realize an identity-isomorphic geometry, whereas a modest-but-reliable CKA indicates weaker structural coupling.

To assess whether the observed CKA exceeds what would be expected by chance, we run a 500-trial permutation test by randomly permuting reader identities to obtain a null distribution.

Identity recoverability (probe). Separately, we train a simple multi-class reader-ID probe (logistic regression) to predict reader IDs from $\mathbf{z}_r$. This probe assesses *decodability*: even when global geometry is not identity-isomorphic, reader identities may still be *separable* in $\mathbf{Z}$ and therefore recoverable by a discriminative classifier. We report probe accuracy and the distribution of prediction confidence (maximum softmax probability).

Results and Interpretation

Structural coupling is statistically reliable but not identity-isomorphic. The observed linear CKA between $\mathbf{Z}$ and $\mathbf{U}$ is 0.116, significantly exceeding the permutation-based null distribution (p=0.002). As shown in Figure 3, CKA under random conditions concentrates around 0.112–0.113, whereas the true value is consistently shifted to the right. Importantly, the absolute CKA magnitude remains modest, suggesting that background variables are *related* to reader identity but do not reproduce a geometry equivalent to a "perfect identity fingerprint" in one-hot space. In other words, the coupling is significant yet *non-isomorphic*.

Reader identity is highly decodable from background variables. Despite the modest global structural alignment, the probe can recover reader IDs from $\mathbf{z}_r$ with 88.9% accuracy (random baseline $\approx 0.34\%$), and confidence is generally high (Figure 4). This demonstrates strong *identity recoverability* from background variables alone. The coexistence of modest CKA and high probe accuracy is expected: CKA evaluates *global* similarity geometry, whereas a discriminative probe can exploit *localized separability* and decision boundaries, enabling high decodability even without identity-isomorphic structure.

Implications for split-dependent gains. Taken together, these findings indicate that background variables contain substantial reader-specific signals. Under splits where reader identities overlap between training and testing (e.g., TEXT), models may leverage these signals as non-transferable shortcuts, yielding memorization-like performance gains. In contrast, in cross-reader settings (e.g., READER and RXT), identity-related cues cannot transfer to unseen readers and may act as confounds that provide little benefit or even harm generalization. Accordingly, we interpret the lack of consistent improvements from background variables in cross-reader scenarios as evidence of *identity-related signal confounding*, rather than the model capturing stable, transferable cognitive characteristics.

9.4 Further Analysis of Absolute R^2

To complement the budget–response analysis in Sec. 4.3, we report *absolute* R^2 with 95% confidence intervals (CIs) for each *non-zero* calibration

budget under both READER and RXT splits (Table 6), together with the corresponding ΔR^2 (Table 7) relative to the uncalibrated model. Overall, the absolute R^2 values mirror the sample-efficiency profile: k-shot residual calibration yields strong early gains for most temporal and count measures, followed by diminishing returns as the calibration budget increases.

READER (cross-reader). With modest calibration, core duration/count measures improve sharply and then saturate. For instance, FFD reaches R^2=0.068 at 0.5% and 0.081 at 1%, stabilizing near 0.093–0.095 by 5–10%; its ΔR^2 is already significant at 0.5% (+0.043 [0.042, 0.044]) and grows smoothly to +0.070 [0.068, 0.073] at 10%. FRD shows a similar front-loaded pattern: R^2 increases to 0.208 (0.5%) and 0.217 (1%), approaching 0.228 at 10%, with ΔR^2 from +0.032 [0.031, 0.033] to +0.052 [0.050, 0.054]. DWL, SKIP, and RUN also rise early and then exhibit diminishing returns, with statistically significant improvements across budgets.

Two indicators are comparatively less sensitive. RPD improves only modestly (e.g., R^2 from 0.036 at 0.5% to $\approx$ 0.042 at 5–10%), albeit with small but significant gains. FIX% is both strong and stable in absolute terms ($R^2 \approx$ 0.382–0.384 with tight CIs across budgets); its marginal changes are near zero at three-decimal resolution, and the sign/significance depends on unrounded values (Table 7).

RXT (cross-reader×cross-text). Under text shift, several measures start lower in absolute terms but exhibit substantial headroom with calibration. FFD increases from R^2=0.020 (0.5%) to 0.034 (1%) and reaches 0.073 at 10%, with significant ΔR^2 throughout (+0.015 [0.014, 0.016] to +0.068 [0.064, 0.072]). FRD and DWL increase steadily (0.129 → 0.172 and 0.208 → 0.250), again with consistently significant gains. SKIP exhibits one of the largest lifts, reaching R^2=0.109 at 10% (with ΔR^2=+0.067 [0.062, 0.071]), and RUN improves similarly (0.084→0.112).

Compared to temporal/count measures, strategy/integration indicators are smaller in magnitude but remain responsive. FRFC rises from R^2=0.178 to 0.210, and RIC increases monotonically (0.046 → 0.067), both with significant improvements across budgets. RPD is borderline at 0.5% (+0.001 with CI touching zero) but becomes reliably positive from 1% onward and reaches +0.013 [0.011, 0.016] at 10%. Finally, FIX% remains stable in absolute terms ($R^2 \approx$ 0.221–0.227), with small but consistently positive ΔR^2 in this setting.

Summary. Across both splits, absolute R^2 and CI-aware ΔR^2 provide a consistent picture: residual calibration yields robust early gains for most duration and count measures at $k \leq 1\%$, followed by diminishing returns by $k \geq 5\%$ for many indicators. FIX% is highly predictable with near-zero marginal change in READER and small positive marginal gains in RXT, while RPD improves modestly and is the least sample-efficient, especially at the smallest calibration budgets.

Table 6: Absolute R^2 with 95% CI for READER and RXT. Color shows positive (green) or negative (red) values. † marks confidence intervals excluding zero.

Ind.	0.5%	1%	2%	5%	10%
READER split (absolute R^2 with 95% CI)					
FFD	+0.068 [0.066, 0.070]†	+0.081 [0.079, 0.083]†	+0.088 [0.085, 0.091]†	+0.093 [0.091, 0.096]†	+0.095 [0.092, 0.098]†
FRD	+0.208 [0.203, 0.214]†	+0.217 [0.212, 0.221]†	+0.224 [0.219, 0.229]†	+0.227 [0.222, 0.232]†	+0.228 [0.223, 0.234]†
FRFC	+0.253 [0.247, 0.259]†	+0.260 [0.254, 0.266]†	+0.264 [0.258, 0.269]†	+0.267 [0.262, 0.273]†	+0.268 [0.262, 0.274]†
RIC	+0.087 [0.082, 0.091]†	+0.092 [0.088, 0.096]†	+0.096 [0.091, 0.101]†	+0.097 [0.093, 0.102]†	+0.098 [0.093, 0.103]†
RPD	+0.036 [0.033, 0.040]†	+0.039 [0.035, 0.043]†	+0.040 [0.036, 0.044]†	+0.042 [0.038, 0.046]†	+0.042 [0.038, 0.046]†
DWL	+0.281 [0.276, 0.286]†	+0.288 [0.284, 0.293]†	+0.294 [0.290, 0.299]†	+0.298 [0.293, 0.302]†	+0.298 [0.294, 0.303]†
FIX%	+0.382 [0.377, 0.388]†	+0.382 [0.377, 0.388]†	+0.383 [0.377, 0.388]†	+0.383 [0.377, 0.389]†	+0.384 [0.378, 0.389]†
SKIP	+0.103 [0.101, 0.106]†	+0.115 [0.113, 0.118]†	+0.121 [0.118, 0.124]†	+0.125 [0.122, 0.128]†	+0.126 [0.123, 0.129]†
RUN	+0.114 [0.111, 0.117]†	+0.118 [0.115, 0.121]†	+0.124 [0.121, 0.127]†	+0.126 [0.123, 0.130]†	+0.127 [0.123, 0.130]†
RXT split (absolute R^2 with 95% CI)					
FFD	+0.020 [0.018, 0.023]†	+0.034 [0.032, 0.037]†	+0.044 [0.041, 0.047]†	+0.068 [0.064, 0.073]†	+0.073 [0.068, 0.078]†
FRD	+0.129 [0.121, 0.138]†	+0.141 [0.132, 0.150]†	+0.150 [0.141, 0.158]†	+0.164 [0.155, 0.172]†	+0.172 [0.163, 0.181]†
FRFC	+0.178 [0.169, 0.188]†	+0.185 [0.175, 0.194]†	+0.193 [0.183, 0.202]†	+0.204 [0.195, 0.214]†	+0.210 [0.200, 0.220]†
RIC	+0.046 [0.039, 0.053]†	+0.050 [0.043, 0.058]†	+0.055 [0.048, 0.062]†	+0.065 [0.057, 0.072]†	+0.067 [0.060, 0.075]†
RPD	+0.017 [0.013, 0.022]†	+0.022 [0.017, 0.026]†	+0.024 [0.020, 0.029]†	+0.028 [0.024, 0.034]†	+0.030 [0.025, 0.036]†
DWL	+0.208 [0.199, 0.219]†	+0.218 [0.208, 0.227]†	+0.229 [0.220, 0.239]†	+0.241 [0.231, 0.250]†	+0.250 [0.241, 0.261]†
FIX%	+0.221 [0.209, 0.233]†	+0.222 [0.211, 0.234]†	+0.223 [0.211, 0.234]†	+0.226 [0.215, 0.237]†	+0.227 [0.216, 0.239]†
SKIP	+0.056 [0.050, 0.062]†	+0.069 [0.063, 0.074]†	+0.085 [0.079, 0.091]†	+0.102 [0.096, 0.108]†	+0.109 [0.103, 0.115]†
RUN	+0.084 [0.078, 0.090]†	+0.086 [0.081, 0.092]†	+0.097 [0.091, 0.103]†	+0.108 [0.101, 0.114]†	+0.112 [0.106, 0.119]†

Table 7: Sample-efficiency of k-shot personalization. Each cell reports the improvement in R^2 after k-shot residual calibration relative to the uncalibrated model, at calibration budgets of 0.5%, 1%, 2%, 5%, and 10% of the target reader's samples. Colors indicate positive (green) or negative (red) change.

Ind.	Δ0.5%	Δ1%	Δ2%	Δ5%	Δ10%
cross-reader (READER)					
FFD	+0.043 [0.042, 0.044]†	+0.056 [0.055, 0.058]†	+0.063 [0.062, 0.065]†	+0.069 [0.066, 0.071]†	+0.070 [0.068, 0.073]†
FRD	+0.032 [0.031, 0.033]†	+0.041 [0.039, 0.042]†	+0.048 [0.046, 0.049]†	+0.051 [0.049, 0.053]†	+0.052 [0.050, 0.054]†
FRFC	+0.021 [0.020, 0.021]†	+0.027 [0.026, 0.029]†	+0.031 [0.030, 0.032]†	+0.035 [0.033, 0.037]†	+0.035 [0.034, 0.037]†
RIC	+0.015 [0.014, 0.015]†	+0.020 [0.019, 0.021]†	+0.024 [0.022, 0.025]†	+0.025 [0.024, 0.026]†	+0.026 [0.024, 0.028]†
RPD	+0.005 [0.005, 0.006]†	+0.008 [0.007, 0.009]†	+0.009 [0.008, 0.010]†	+0.011 [0.010, 0.012]†	+0.011 [0.010, 0.012]†
DWL	+0.037 [0.036, 0.038]†	+0.044 [0.043, 0.045]†	+0.050 [0.049, 0.052]†	+0.053 [0.052, 0.055]†	+0.054 [0.052, 0.056]†
FIX%	-0.000 [-0.001, -0.000]†	-0.000 [-0.000, 0.000]	+0.000 [0.000, 0.001]†	+0.001 [0.001, 0.001]†	+0.001 [0.001, 0.001]†
SKIP	+0.032 [0.031, 0.033]†	+0.044 [0.043, 0.046]†	+0.050 [0.048, 0.051]†	+0.054 [0.052, 0.055]†	+0.055 [0.053, 0.057]†
RUN	+0.025 [0.024, 0.026]†	+0.029 [0.028, 0.030]†	+0.034 [0.033, 0.036]†	+0.037 [0.035, 0.038]†	+0.037 [0.035, 0.039]†
cross-reader×cross-text (RXT)					
FFD	+0.015 [0.014, 0.016]†	+0.029 [0.027, 0.031]†	+0.039 [0.037, 0.041]†	+0.063 [0.059, 0.067]†	+0.068 [0.064, 0.072]†
FRD	+0.011 [0.010, 0.012]†	+0.023 [0.021, 0.024]†	+0.032 [0.030, 0.034]†	+0.046 [0.043, 0.048]†	+0.054 [0.050, 0.058]†
FRFC	+0.008 [0.007, 0.008]†	+0.014 [0.013, 0.015]†	+0.022 [0.020, 0.024]†	+0.033 [0.031, 0.036]†	+0.039 [0.036, 0.043]†
RIC	+0.004 [0.004, 0.005]†	+0.008 [0.007, 0.009]†	+0.014 [0.012, 0.015]†	+0.023 [0.021, 0.025]†	+0.026 [0.023, 0.029]†
RPD	+0.001 [-0.000, 0.001]	+0.005 [0.004, 0.006]†	+0.007 [0.006, 0.009]†	+0.012 [0.010, 0.014]†	+0.013 [0.011, 0.016]†
DWL	+0.012 [0.012, 0.013]†	+0.021 [0.020, 0.023]†	+0.033 [0.031, 0.035]†	+0.045 [0.042, 0.048]†	+0.054 [0.051, 0.058]†
FIX%	+0.001 [0.001, 0.001]†	+0.002 [0.001, 0.002]†	+0.003 [0.002, 0.004]†	+0.006 [0.005, 0.007]†	+0.007 [0.006, 0.008]†
SKIP	+0.014 [0.013, 0.015]†	+0.027 [0.025, 0.028]†	+0.043 [0.040, 0.045]†	+0.060 [0.056, 0.064]†	+0.067 [0.062, 0.071]†
RUN	+0.012 [0.011, 0.013]†	+0.014 [0.012, 0.015]†	+0.024 [0.022, 0.026]†	+0.035 [0.032, 0.038]†	+0.039 [0.036, 0.043]†

Table 8: Eye-tracking indicators with Δall and Δperm (R^2 with CI95). Each cell shows R^2 and CI95 in separate columns. Five fields per domain: w, all, Δall, perm, Δperm.

Indicator	Model	READER										RXT										TEXT									
		w		all		Δall		perm		Δperm		w		all		Δall		perm		Δperm		w		all		Δall		perm		Δperm	
		R2	CI95	R2	CI95	R2	CI95	R2	CI95	R2	CI95	R2	CI95	R2	CI95	R2	CI95	R2	CI95	R2	CI95	R2	CI95	R2	CI95	R2	CI95	R2	CI95	R2	CI95
Early-stage (initial lexical processing)																															
FFD	ElasticNet	0.008	[0.007,0.010]	-0.003	[-0.004,-0.001]	-0.011	[-0.012,-0.010]	0.008	[0.007,0.010]	-0.001	[-0.002,-0.001]	0.005	[0.003,0.008]	-0.004	[-0.008,-0.000]	-0.009	[-0.012,-0.007]	-0.034	[-0.039,-0.028]	-0.039	[-0.044,-0.034]	0.009	[0.008,0.010]	0.014	[0.013,0.016]	0.005	[0.004,0.006]	0.015	[0.014,0.016]	0.006	[0.005,0.007]
	LightGBM	0.025	[0.023,0.027]	0.004	[0.002,0.006]	-0.021	[-0.022,-0.019]	-0.004	[-0.006,-0.002]	-0.029	[-0.031,-0.027]	0.005	[0.003,0.008]	-0.011	[-0.015,-0.006]	-0.016	[-0.019,-0.012]	-0.015	[-0.020,-0.011]	-0.021	[-0.025,-0.017]	0.009	[0.008,0.010]	0.080	[0.077,0.083]	0.070	[0.068,0.073]	0.080	[0.077,0.083]	0.071	[0.068,0.073]
	Transformer	0.022	[0.020,0.025]	-0.051	[-0.054,-0.047]	-0.073	[-0.076,-0.069]	-0.015	[-0.018,-0.011]	-0.037	[-0.040,-0.034]	-0.014	[-0.019,-0.010]	-0.103	[-0.113,-0.094]	-0.089	[-0.098,-0.079]	-0.077	[-0.087,-0.068]	-0.063	[-0.071,-0.054]	-0.007	[-0.009,-0.005]	0.070	[0.066,0.074]	0.077	[0.073,0.081]	0.067	[0.063,0.071]	0.074	[0.071,0.078]
FRD	ElasticNet	0.136	[0.132,0.140]	0.129	[0.125,0.134]	-0.007	[-0.008,-0.006]	0.124	[0.120,0.129]	-0.012	[-0.013,-0.011]	0.123	[0.115,0.131]	0.121	[0.113,0.129]	-0.002	[-0.005,0.000]	0.123	[0.116,0.131]	-0.000	[-0.002,0.002]	0.116	[0.112,0.120]	0.130	[0.126,0.133]	0.014	[0.013,0.015]	0.121	[0.117,0.124]	0.005	[0.004,0.006]
	LightGBM	0.176	[0.171,0.181]	0.145	[0.140,0.151]	-0.031	[-0.034,-0.029]	0.147	[0.143,0.151]	-0.029	[-0.032,-0.027]	0.118	[0.110,0.127]	0.102	[0.091,0.112]	-0.016	[-0.021,-0.012]	0.071	[0.060,0.083]	-0.047	[-0.054,-0.041]	0.114	[0.110,0.118]	0.186	[0.181,0.191]	0.072	[0.069,0.074]	0.185	[0.180,0.190]	0.071	[0.068,0.073]
	Transformer	0.186	[0.180,0.191]	0.063	[0.056,0.071]	-0.122	[-0.128,-0.117]	0.043	[0.035,0.051]	-0.143	[-0.150,-0.136]	0.058	[0.046,0.069]	-0.003	[-0.021,0.013]	-0.061	[-0.076,-0.047]	-0.013	[-0.029,0.003]	-0.071	[-0.086,-0.057]	0.054	[0.049,0.059]	0.143	[0.137,0.149]	0.089	[0.084,0.094]	0.159	[0.152,0.165]	0.104	[0.100,0.109]
FRFC	Poisson	0.183	[0.177,0.188]	0.193	[0.188,0.198]	0.010	[0.009,0.011]	0.178	[0.172,0.183]	-0.005	[-0.006,-0.004]	0.172	[0.162,0.182]	0.186	[0.176,0.197]	0.014	[0.011,0.017]	-0.047	[-0.078,-0.020]	-0.219	[-0.248,-0.193]	0.161	[0.156,0.166]	0.176	[0.171,0.181]	0.015	[0.013,0.016]	0.165	[0.161,0.170]	0.004	[0.003,0.005]
	LightGBM	0.210	[0.205,0.215]	0.199	[0.193,0.204]	-0.011	[-0.013,-0.009]	0.163	[0.159,0.167]	-0.047	[-0.050,-0.045]	0.148	[0.140,0.156]	0.154	[0.145,0.164]	0.006	[0.003,0.010]	0.137	[0.127,0.146]	-0.011	[-0.014,-0.008]	0.145	[0.141,0.149]	0.199	[0.195,0.203]	0.054	[0.052,0.056]	0.200	[0.196,0.204]	0.055	[0.053,0.057]
	Transformer	0.244	[0.238,0.250]	0.166	[0.160,0.172]	-0.077	[-0.081,-0.072]	0.169	[0.164,0.174]	-0.074	[-0.078,-0.071]	0.138	[0.128,0.148]	0.150	[0.139,0.161]	0.012	[0.004,0.020]	0.082	[0.071,0.093]	-0.056	[-0.065,-0.047]	0.115	[0.110,0.120]	0.195	[0.189,0.200]	0.079	[0.076,0.083]	0.194	[0.189,0.199]	0.078	[0.075,0.082]
Late-stage (regression / integration processing)																															
RIC	Poisson	0.030	[0.027,0.033]	0.031	[0.028,0.035]	0.001	[0.000,0.002]	0.026	[0.022,0.029]	-0.005	[-0.006,-0.004]	0.032	[0.025,0.039]	0.033	[0.026,0.040]	0.001	[-0.000,0.003]	0.028	[0.022,0.035]	-0.004	[-0.005,-0.002]	0.038	[0.035,0.041]	0.040	[0.037,0.043]	0.002	[0.001,0.003]	0.041	[0.038,0.044]	0.003	[0.002,0.004]
	LightGBM	0.052	[0.048,0.056]	0.035	[0.031,0.039]	-0.016	[-0.018,-0.015]	0.042	[0.038,0.047]	-0.009	[-0.011,-0.008]	0.017	[0.011,0.023]	0.017	[0.010,0.025]	0.001	[-0.003,0.005]	0.016	[0.009,0.023]	-0.000	[-0.003,0.002]	0.023	[0.020,0.026]	0.059	[0.055,0.063]	0.036	[0.034,0.038]	0.057	[0.053,0.061]	0.034	[0.032,0.036]
	Transformer	0.077	[0.072,0.081]	-0.021	[-0.026,-0.016]	-0.098	[-0.102,-0.094]	-0.021	[-0.025,-0.016]	-0.097	[-0.101,-0.093]	0.014	[0.008,0.021]	-0.071	[-0.085,-0.057]	-0.086	[-0.097,-0.072]	-0.110	[-0.126,-0.093]	-0.124	[-0.139,-0.111]	0.019	[0.015,0.022]	0.042	[0.037,0.047]	0.023	[0.020,0.026]	0.046	[0.042,0.050]	0.028	[0.024,0.030]
RPD	ElasticNet	0.020	[0.018,0.022]	0.019	[0.017,0.021]	-0.001	[-0.002,-0.000]	0.017	[0.015,0.019]	-0.002	[-0.003,-0.001]	0.020	[0.018,0.023]	0.021	[0.018,0.023]	0.001	[0.000,0.002]	0.020	[0.017,0.022]	0.000	[-0.001,0.001]	0.018	[0.016,0.020]	0.022	[0.020,0.024]	0.004	[0.003,0.005]	0.020	[0.018,0.023]	0.002	[0.001,0.002]
	LightGBM	0.031	[0.029,0.033]	0.019	[0.017,0.021]	-0.012	[-0.013,-0.010]	0.016	[0.014,0.018]	-0.015	[-0.017,-0.014]	0.017	[0.014,0.019]	0.007	[0.003,0.013]	-0.010	[-0.013,-0.007]	0.017	[0.010,0.025]	0.000	[-0.003,0.002]	0.015	[0.013,0.017]	0.031	[0.029,0.033]	0.016	[0.015,0.019]	0.031	[0.029,0.032]	0.016	[0.015,0.018]
	Transformer	0.022	[0.016,0.028]	-0.055	[-0.062,-0.048]	-0.077	[-0.086,-0.068]	-0.008	[-0.012,-0.005]	-0.030	[-0.037,-0.024]	-0.008	[-0.017,-0.000]	-0.507	[-0.615,-0.418]	-0.500	[-0.604,-0.418]	-0.043	[-0.064,-0.026]	-0.036	[-0.055,-0.019]	-0.000	[-0.004,0.004]	-0.006	[-0.013,0.002]	-0.006	[-0.012,0.001]	0.020	[0.016,0.024]	0.020	[0.016,0.024]
DWELL	ElasticNet	0.206	[0.202,0.210]	0.203	[0.199,0.207]	-0.003	[-0.005,-0.002]	0.205	[0.201,0.209]	-0.001	[-0.002,-0.001]	0.200	[0.191,0.208]	0.204	[0.195,0.213]	0.004	[0.001,0.007]	0.195	[0.186,0.203]	-0.005	[-0.007,-0.003]	0.185	[0.181,0.189]	0.207	[0.203,0.211]	0.022	[0.021,0.023]	0.189	[0.185,0.193]	0.004	[0.004,0.005]
	LightGBM	0.244	[0.239,0.249]	0.216	[0.210,0.221]	-0.028	[-0.031,-0.026]	0.205	[0.201,0.209]	-0.039	[-0.041,-0.037]	0.196	[0.187,0.205]	0.179	[0.168,0.190]	-0.017	[-0.022,-0.012]	0.175	[0.164,0.186]	-0.021	[-0.025,-0.016]	0.187	[0.183,0.191]	0.286	[0.281,0.291]	0.099	[0.095,0.102]	0.284	[0.280,0.289]	0.097	[0.094,0.100]
	Transformer	0.259	[0.254,0.264]	0.004	[-0.005,0.012]	-0.255	[-0.262,-0.248]	0.122	[0.115,0.130]	-0.136	[-0.143,-0.131]	0.184	[0.174,0.193]	-0.154	[-0.180,-0.129]	-0.338	[-0.361,-0.315]	-0.081	[-0.102,-0.061]	-0.264	[-0.282,-0.247]	0.165	[0.161,0.170]	0.252	[0.247,0.258]	0.087	[0.083,0.092]	0.242	[0.235,0.248]	0.077	[0.072,0.081]
Global (overall reading behavior)																															
FIX%	ElasticNet	0.217	[0.213,0.221]	0.217	[0.213,0.221]	0.000	[0.000,0.000]	0.217	[0.213,0.221]	0.000	[0.000,0.000]	0.225	[0.217,0.234]	0.225	[0.217,0.233]	0.000	[0.000,0.000]	0.225	[0.217,0.233]	0.000	[0.000,0.000]	0.213	[0.209,0.216]	0.213	[0.209,0.216]	0.000	[0.000,0.000]	0.213	[0.209,0.216]	0.000	[0.000,0.000]
	LightGBM	0.383	[0.377,0.388]	0.374	[0.368,0.379]	-0.009	[-0.010,-0.008]	0.374	[0.368,0.379]	-0.009	[-0.010,-0.008]	0.218	[0.206,0.229]	0.216	[0.205,0.227]	-0.002	[-0.003,-0.001]	0.214	[0.202,0.225]	-0.004	[-0.005,-0.003]	0.209	[0.205,0.214]	0.211	[0.206,0.216]	0.002	[0.001,0.002]	0.210	[0.206,0.215]	0.001	[0.001,0.002]
	Transformer	0.457	[0.452,0.462]	0.423	[0.417,0.429]	-0.034	[-0.037,-0.032]	0.427	[0.422,0.433]	-0.030	[-0.032,-0.028]	0.377	[0.366,0.387]	0.385	[0.374,0.396]	0.009	[0.004,0.013]	0.380	[0.368,0.391]	0.004	[-0.002,0.009]	0.349	[0.345,0.354]	0.369	[0.364,0.374]	0.020	[0.017,0.022]	0.365	[0.360,0.369]	0.015	[0.013,0.017]
SKIP	ElasticNet	0.049	[0.047,0.052]	0.047	[0.044,0.049]	-0.003	[-0.004,-0.002]	-0.013	[-0.018,-0.009]	-0.063	[-0.067,-0.059]	0.041	[0.035,0.046]	0.039	[0.032,0.046]	-0.001	[-0.004,0.001]	0.026	[0.020,0.033]	-0.014	[-0.017,-0.011]	0.051	[0.048,0.053]	0.056	[0.053,0.058]	0.005	[0.004,0.006]	0.055	[0.052,0.057]	0.004	[0.004,0.005]
	LightGBM	0.071	[0.069,0.074]	0.045	[0.042,0.048]	-0.026	[-0.028,-0.025]	0.042	[0.039,0.044]	-0.030	[-0.031,-0.028]	0.042	[0.036,0.048]	0.023	[0.015,0.030]	-0.019	[-0.023,-0.016]	0.028	[0.020,0.035]	-0.015	[-0.018,-0.012]	0.052	[0.050,0.054]	0.116	[0.113,0.119]	0.064	[0.062,0.066]	0.116	[0.112,0.118]	0.064	[0.062,0.066]
	TransformerStable	0.092	[0.089,0.095]	0.057	[0.054,0.060]	-0.035	[-0.037,-0.033]	0.063	[0.060,0.066]	-0.029	[-0.031,-0.027]	0.037	[0.030,0.043]	-0.060	[-0.072,-0.049]	-0.097	[-0.104,-0.089]	-0.015	[-0.023,-0.007]	-0.051	[-0.056,-0.047]	0.050	[0.047,0.052]	0.055	[0.051,0.058]	0.005	[0.003,0.008]	0.047	[0.044,0.050]	-0.003	[-0.005,-0.001]
RUN	Poisson	0.064	[0.061,0.067]	0.062	[0.059,0.065]	-0.002	[-0.003,-0.001]	0.061	[0.058,0.064]	-0.003	[-0.004,-0.002]	0.068	[0.061,0.074]	0.070	[0.063,0.077]	0.002	[-0.000,0.005]	0.049	[0.041,0.056]	-0.019	[-0.022,-0.016]	0.071	[0.068,0.074]	0.082	[0.079,0.085]	0.011	[0.010,0.012]	0.071	[0.069,0.074]	0.000	[0.000,0.001]
	LightGBM	0.055	[0.051,0.059]	0.035	[0.031,0.039]	-0.020	[-0.022,-0.019]	0.032	[0.029,0.035]	-0.023	[-0.025,-0.021]	0.031	[0.024,0.039]	0.028	[0.019,0.036]	-0.004	[-0.007,-0.001]	0.018	[0.010,0.027]	-0.013	[-0.015,-0.010]	0.044	[0.041,0.047]	0.109	[0.105,0.113]	0.065	[0.063,0.067]	0.108	[0.104,0.112]	0.064	[0.062,0.066]
	Transformer	0.081	[0.078,0.085]	-0.034	[-0.038,-0.029]	-0.115	[-0.118,-0.112]	-0.036	[-0.040,-0.031]	-0.117	[-0.121,-0.113]	0.061	[0.053,0.069]	-0.109	[-0.123,-0.097]	-0.170	[-0.181,-0.160]	-0.058	[-0.070,-0.046]	-0.119	[-0.130,-0.109]	0.050	[0.046,0.054]	0.111	[0.106,0.115]	0.060	[0.058,0.063]	0.112	[0.109,0.116]	0.062	[0.059,0.065]