

A Systematic Analysis of the Predictive Power of LM Surprisal in Reading Chinese

Hongao Zhu^{♣*}, Muxiaoqiao Xu^{†*}, Yikang Liu[‡], Siyuan Song^{†◇},
Yuxia Wang[†], Byung-Doh Oh[‡], Hai Hu[♡]

[♣]Department of Linguistics, UC San Diego [†]School of Foreign Languages, Shanghai Jiao Tong University

[‡]School of Computer Science, Shanghai Jiao Tong University [◇]Department of Linguistics, UT Austin

[‡]Division of Linguistics and Multilingual Studies, Nanyang Technological University

[♡]Dept. of Language Science and Technology/Division of AI and the Humanities, Hong Kong Polytechnic University

*Equal contribution. Correspondence to hai.hu@polyu.edu.hk

Abstract

This study analyzes the predictive power of LM-derived, token-level surprisal on Mandarin Chinese reading times. We first propose the Shortest Matching Sequence (SMS), an alignment scheme that maps between the word segmentation assumed by eye-tracking corpora and the LMs’ subword tokenization, as the two tokenizations often disagree in the context of Mandarin Chinese. Then, using a suite of Chinese-Pythia models (14M–1.4B) trained on scratch with 30B tokens, we examine how well surprisal predicts first fixation duration, gaze duration, and total reading time in three paragraph-level eye-tracking corpora of Mandarin Chinese (GECO-CN, HKP, and MECO). Contrary to previous null findings, our results show that surprisal is predictive of Chinese reading times. However, whether predictive power scales with model size and the amount of training is corpus-specific: bigger models predict better in GECO-CN, whereas inverse scaling emerges in HKP and, at the largest sizes, in MECO. Subsequently, we tested one possible explanation for the inverse scaling in HKP and found that checkpoints whose surprisal remains closer to n -gram statistics are better predictors of reading. All in all, the predictive power of surprisal on Chinese reading time measurements is corpus-specific, which cautions against drawing scaling conclusions from a single corpus.

1 Introduction

Humans read text incrementally, allocating different time and effort for each word they encounter. The exact time and effort spent for each word has been proposed to be affected by the word’s predictability. This proposal has been formalized by surprisal theory (Hale, 2001; Levy, 2008), which states that the more unexpected a word is, the higher the effort required to process it. Such unexpectedness can be quantified by the probability of the word in question given its preceding context.

Many studies have investigated whether surprisal estimated from Large Language Models (LLMs) can predict reading times obtained from human subjects in reading English text (Wilcox et al., 2020; Shain et al., 2022; Oh and Schuler, 2023b; Pimentel et al., 2023; Kuribayashi et al., 2025, among others). Beyond English, several studies have examined LM-derived surprisal in other languages, such as Japanese (Kuribayashi et al., 2021, 2022), Dutch (Meister et al., 2021; Boeve and Bogaerts, 2025), or multiple languages together (Xu et al., 2023; Wilcox et al., 2023b; De Varda and Marelli, 2023).

Unlike studies in English, which often make use of multiple eye-tracking datasets, most studies in other languages only utilized one dataset, and are thus limited in the generalizability of their findings (Kuribayashi et al., 2021, 2022). For Mandarin Chinese, only two studies have examined whether surprisal estimated from LLMs can predict human reading times, and both reported negative results (Xu et al., 2023; Sun and Liu, 2025).¹ In contrast, positive results have been obtained from almost all other languages (Wilcox et al., 2023b; Xu et al., 2023). Therefore, it remains an open question whether and why Mandarin Chinese diverges from established cross-linguistic patterns.

Furthermore, how language models’ scale and quality correlate with the predictive power of their surprisal is poorly understood outside English (Oh and Schuler, 2023a,b; Oh et al., 2024). Systematic examinations of scaling effects in non-English languages have been constrained by the lack of controlled, monolingual LLM suites like Pythia (Biderman et al., 2023) or OPT (Zhang et al., 2022b): Wilcox et al. (2023b) only evaluated LMs with parameter sizes of 30M and 1.3B, which they have trained from scratch; De Varda and Marelli (2023)

¹One older work (Fan and Reilly, 2020) used surprisal estimates from Chinese n -gram models to predict reading time in Chinese children.

examined XGLM models from 564M to 4.5B parameters, which were trained on multilingual data (Lin et al., 2022); and Boeve and Bogaerts (2025) tested models which differed in architecture and training data, and did not include intermediate checkpoints through the course of training.

A parallel limitation lies in the scope of the eye-tracking corpus: most non-English studies have relied on a single, small dataset (Wilcox et al., 2023b; Xu et al., 2023; De Varda and Marelli, 2023).

To make our study more generalizable, we overcome the above limitations by testing surprisal estimates across three large-scale eye-tracking corpora (GECO-CN, HKP, and MECO; Sui et al., 2023; Wu and Kit, 2023; Siegelman et al., 2022), spanning narrative fiction, news reports, and encyclopedic prose. This study asks the following research questions, with a focus on Mandarin Chinese:

1. To what extent is surprisal estimated from LMs predictive of Chinese reading times?
2. Does the scale of training data or model size have an inverse relationship with predictive power of surprisal?
3. How can the corpus-specific scaling of surprisal’s predictive power across models be explained?

To address these questions, we analyze First Fixation Duration (FFD), Gaze Duration (GD), and Total Reading Time (TRT) across Mandarin eye-tracking corpus using linear mixed-effects regressions. We analyze these behavioral data against a Chinese-Pythia model suite with five parameter sizes (14M — 1.4B) across training.

Our contributions are: (1) an alignment scheme for differing Chinese tokenizations across eye-tracking corpora and LMs, (2) a suite of Chinese decoder-only LMs trained from scratch on 30B tokens for psycholinguistic evaluation, (3) the first comprehensive demonstration that LM surprisal predicts Chinese reading times, while showing that model-size and training-volume scaling are corpus-specific, (4) evidence that the surprisal-to-reading-time relationship varies across corpora, cautioning against drawing scaling conclusions from a single corpus, and (5) evidence from HKP that checkpoints remaining closest to local n -gram statistics best predict reading times, offering a candidate mechanism for its post-peak decline.

2 Related Work

2.1 Surprisal and reading times in Chinese

There is a lot of work in English on using information-theoretic metrics extracted from LMs to predict human reading time (Oh et al., 2022; Oh and Schuler, 2023b; Pimentel et al., 2023; Shain et al., 2024). Although this line of inquiry has also expanded to other languages like Japanese (Kuribayashi et al., 2021, 2022) and Dutch (Meister et al., 2021; Boeve and Bogaerts, 2025), Mandarin Chinese remains comparatively understudied. The two existing studies found null results for surprisal’s predictive power of Mandarin Chinese reading time from LMs (Xu et al., 2023; Sun and Liu, 2025). Specifically, Xu et al. (2023) studied six languages and found that surprisal computed from mGPT and GPT2 does not add predictive power for eye-tracking measures in reading Mandarin Chinese. Sun and Liu (2025) devised a new attention-based metric and compared it against other existing metrics including surprisal; they reported that word surprisal is not a significant predictor in their Generalized Additive Mixed Models (GAMMs; Wood, 2025), yielding only a weak effect on eye movements.

It is thus an open question whether Mandarin Chinese poses a unique challenge for using surprisal to predict human eye-tracking measurements. Besides, Mandarin Chinese, among other languages, presents a challenge in the mismatch between the tokenization from LMs and the tokenization in eye-tracking corpus.

2.2 Inverse scaling of surprisal’s predictive power

While the predictive power of surprisal in English has been widely replicated, an emerging line of work reports a counterintuitive inverse relationship between model size and the ability of surprisal to predict human reading times (Oh and Schuler, 2023b; Oh et al., 2024). Several explanations have been proposed for this “bigger is not always better” phenomenon (Wilcox et al., 2025). First, larger models appear to over-predict rare words: they learn superhumanly complex associations for low-frequency lexical items, which diverge from human expectations and thereby reduce the correlation with reading times (Oh et al., 2024). Second, the scaling mismatch may be driven by systematic biases tied to word class: for instance, larger models tend to under-predict reading times for named enti-

ties while over-predicting them for function words, introducing a distortion that grows with model capacity (Oh and Schuler, 2023b). Third, this inverse scaling cannot be attributed to data leakage, as the effect persists even when models are trained on carefully filtered, leakage-free corpora (Oh et al., 2025). Taken together, these findings suggest that as language models scale up, their internal probability distributions become less human-like with respect to reading-time prediction. This raises important questions about the appropriateness of using ever-larger models as cognitive models of human processing.

Our study adds a new dimension to this debate by comparing corpora that differ in their materials: a detective novel (GECO-CN), news reports (HKP), and Wikipedia-style paragraphs (MECO). If the surprisal-to-reading-time relationship depends on the materials, then corpus-specific scaling patterns could emerge. This possibility motivates our cross-corpus comparison and underscores the need to examine scaling effects across diverse linguistic materials.

2.3 Systematic scaling studies beyond English

A major obstacle to investigating the scaling effects is the scarcity of model suites that are trained on identical data with publicly available intermediate training checkpoints. In English, models suites such as Pythia (Biderman et al., 2023) and OPT (Zhang et al., 2022b) has enabled controlled comparisons of surprisal estimates across model sizes and training steps. However, comparable resources hardly exist in other languages, so researchers often rely on heterogeneous models with different architectures, training data, and tokenization schemes, which have been confounding the scaling analyses. In Dutch, for example, Boeve and Bogaerts (2025) compared multiple LMs of varying sizes and architectures, but the models did not share consistent training data, neither were the intermediate checkpoints released. Thus, a clean dissociation of model size from other factors can not be made.

Our work addresses this gap by training a suite of decoder-only models—referred to as Chinese-Pythia—with five parameter sizes (14M to 1.4B) on the same 30B-token Chinese corpus. Because all models share the same architecture, tokenizer, and training data, and because intermediate checkpoints are released, we are able to systematically examine how surprisal’s predictive power scales with model

size and training amount. This allows us to test whether the inverse scaling observed in English generalizes to a typologically distinct language and to explore potential moderators.

3 Experimental Setup

3.1 Eye-tracking data

We select three large-scale eye-tracking datasets for natural Mandarin Chinese reading – the Ghent Eye-tracking COrpus of sentence reading for Chinese-English bilinguals (GECO-CN; Sui et al., 2023), the Hong Kong Corpus of Chinese Sentence and Passage Reading (HKC; Wu and Kit, 2023), and the Multilingual Eye-Movement Corpus (MECO; Siegelman et al., 2025).² GECO-CN records eye movements from native Chinese speakers reading a detective novel translated into simplified Chinese. HKC comprises simplified Chinese texts derived from Xinhua news reports. MECO contains eye-tracking data collected from Wikipedia-style short passages in simplified and traditional Chinese. To ensure comparability across corpora, we included only the simplified-Chinese, passage-reading portion of data from all three sources. Several commonly used Chinese reading time corpus were excluded because they either do not provide all eye-tracking metrics (Pan et al., 2022), or their stimuli are sentence-level materials used in controlled psycholinguistic experiments (Zhang et al., 2022a).

We analyze three standard measurements of reading time (Rayner, 1998): (1) **First Fixation Duration** (FFD), the duration of the initial fixation on an interest area during its first pass; (2) **Gaze Duration** (GD), the sum of all first-pass fixations on the interest area; and (3) **Total Reading Time** (TRT), the cumulative duration of all fixations on the interest area across all passes. Following Smith and Levy (2013) and Xu et al. (2023), we preprocessed each reading-time measure by excluding skipped words (0 ms reading time), extreme outliers (>2000 ms), tokens with non-finite surprisal, and paragraph-initial words. We refrained from applying a hard surprisal cutoff to avoid disproportionately removing multi-token Chinese words.³

Table 1 summarizes the three corpora after filtering. We report perplexity under Qwen2.5-0.5B (Yang et al., 2024), which is a reference model

²Details of the raw corpus are presented in Appendix A.

³Observation counts remaining after each filtering step are listed in Table 5

Corpus	N_{subj}	N_{words}	PPL	TRT		FFD		GD	
				N_{obs}	ms	N_{obs}	ms	N_{obs}	ms
GECO-CN	30	46,680	60.9	348,105	281.8 (196.3)	348,527	206.5 (89.6)	253,298	221.4 (115.7)
HKP	98	5,071	46.6	195,840	289.7 (202.1)	196,048	209.8 (91.8)	158,274	222.4 (114.1)
MECO	39	1,812	31.5	27,192	358.5 (256.1)	26,293	227.0 (101.9)	26,283	248.5 (135.5)

Table 1: Descriptive statistics of the data entering the regression analyses, i.e., after preprocessing (Appendix A). HKP refers to the passage-reading subset of HKC, and MECO refers to the simplified-Chinese portion. N_{words} is the number of distinct word positions in the text (aligned between corpus and model tokenization; Section 3.4) with at least one retained TRT observation. N_{obs} is the number of subject-word observations retained for each reading-time measure; counts are identical across all model sizes and checkpoints. Reading times are reported as mean (standard deviation) in ms over the retained observations. Perplexity is 2 raised to the mean token-level surprisal (bits) under Qwen2.5-0.5B (Qwen et al., 2025), a fixed reference model independent of the Chinese-Pythia suite under evaluation; as a model-relative quantity, it indexes corpus difficulty only under this reference.

independent of the Chinese-Pythia suite used in the main analysis. Notably, MECO is read most slowly (359 ms mean TRT) despite having the lowest perplexity (31.5), whereas GECO-CN combines the highest perplexity (60.9) with the fastest mean reading times (282 ms).

3.2 Models

We train a series of decoder-only models in Mandarin Chinese, following the Pythia model suite (Biderman et al., 2023), with similar configurations. Our Chinese-Pythia suite includes models with the following parameter size: 14M, 70M, 160M, 410M and 1.4B. We use the tokenizer from Chinese-Llama (Cui et al., 2024). The training data contains 30B tokens and is a subset of the CCI3.0-HQ corpus (Wang et al., 2024), which is a deduplicated collection of Chinese web text covering news, social media, and blogs. We set the context length to be 256 and the number of tokens trained per step is 40,960. For each size, we analyze 40 saved training checkpoints: the 14 early checkpoints saved according to a doubling schedule (after 2^0 to 2^{13} , or 1 to 8,192 training steps), the checkpoint after 10,000 steps, and the 25 checkpoints saved after every 30,000 steps (from 30,000 to 750,000). Together these span from the first training step to the end of the 30B-token run. All model training was done on two A100 GPUs.⁴

3.3 Estimating surprisal from LMs

For surprisal, we follow the literature in using the negative log probability of a token as its surprisal (Oh and Schuler, 2023b,a; Wilcox et al., 2023a; Pimentel et al., 2023; Wilcox et al., 2023b). That is, for a token x_t perceived at index t with preceding

context $x_{<t}$, its surprisal is defined as⁵:

$$S(x_t|x_{<t}) \equiv -\log p(x_t|x_{<t})$$

For a multi-token word $w = \{x_t, x_{t+1}, \dots, x_n\}$, we sum up the negative log probabilities of the tokens:

$$S(w) = \sum_{i=t}^n -\log p(x_i|x_{<i})$$

3.4 Aligning tokenizations between LMs and eye-tracking data in Chinese

Aligning LM tokens with the word segmentation of an eye-tracking corpus is more complicated in Mandarin Chinese than in languages written with spaces, such as English and Dutch. In those languages, LM sub-word tokens almost always nest within corpus words, so the surprisal of a word is simply the sum of its sub-token surprisals. In Mandarin Chinese, the corpus segmentation and the LM tokenizer segment the same character string independently, so a corpus word and an LM token can coincide, the LM tokenization can be finer, the corpus segmentation can be finer, or their boundaries can cross (Table 2; see Table 3 for the frequency of each condition). We therefore align the two sequences into the shortest spans whose boundaries are shared by both tokenizations, and compute surprisal and reading times for each span.

On the LM side, the surprisal of a span is the sum of the surprisals of its LM tokens. When the LM tokenization is finer, this is optimal: the span is a single corpus word, whose reading times are used as they are.

⁴Our models are available at <https://huggingface.co/collections/yikang0131/zh-pythia-v02>.

⁵We modify the scripts from Haller et al. (2024) <https://github.com/DiLi-Lab/LM-cog-profiles> to compute the surprisal in Chinese.

CHINESE HKP eye-tracking corpus + Chinese-LLaMA tokenizer			
Corpus	国内 企业 的 老总 们 也 以 同样 热情 关注 这些 人才		
LM	国内 企业的 老 总 们 也 以 同样 热情 关注 这些人 才		
Aligned	国内 企业的 老总 们 也 以 同样 热情 关注 这些人才		
Gloss	Domestic company-DE leaders PL also with same enthusiasm focus-on these talents		
Trans.	Leaders from domestic companies also focus on these talents with the same enthusiasm.		

Mismatch type	Corpus tokens	LM tokens	Alignment operation
LM finer	老总	老 总	Sum LM surprisals
Corpus finer	企业 的	企业的	Sum TRTs; Take first for FFD; Skip for GD
Crossing	这些 人才	这些人 才	Merge both sides

Table 2: Alignment between eye-tracking corpus tokenization and LM tokenization in Chinese. Token boundaries are shown as “|”. Colors mark the three types of boundary mismatch; each is resolved by aligning to the smallest span on which both tokenizations agree, with the operations listed in the lower panel.

On the eye-tracking side, spans that contain more than one corpus word require combining word-level reading times, and the three measures differ in whether this is possible. TRT, the total time spent on a region across all passes, is additive, so the TRT of a span is the sum of the TRTs of its words. FFD, the duration of the first fixation on the span, equals the FFD of whichever of its words was fixated first, so we take the FFD of that word. GECON and HKP provide the information needed to identify this word, but MECO does not, so multi-word spans are excluded from the FFD analysis of MECO. GD, the sum of first-pass fixations on the span, cannot be reconstructed from the GDs of its words, because it depends on whether the reader left the span between fixating them, which the corpora do not record.⁶ We therefore exclude the GDs of multi-word spans, which make up 21.7% of aligned spans in GECON, 14.1% in HKP, and 10.0% in MECO (See Table 3).

3.5 Regression models

We follow previous work to fit linear mixed-effects models to human reading data, with measures for reading time as the dependent variable, and word length, word frequency, and optionally surprisal as the independent variables. All predictors are standardized, and the predictors of the current word are accompanied by their spillover terms for the

⁶For example, consider a span ABC in which the reader first fixated B and then A. If the reader then left the span before fixating C, $GD(ABC) = GD(B) + GD(A)$; if they moved from A directly to C, $GD(ABC) = GD(B) + GD(A) + GD(C)$. Word-level GDs do not distinguish between the two cases.

preceding word. The baseline model includes word length, log-transformed word frequency, and their spillover terms, with by-subject random slopes for word length; the surprisal model adds standardized surprisal and its spillover term. We compute the difference in log-likelihood (Δllh) between the two models to measure the predictive power of surprisal.⁷ Because Δllh is uniformly positive across all 40 checkpoints for every model size, corpus, and measure (Section 4), we treat its magnitude, rather than per-fit significance, as the unit of comparison.

3.6 Quantifying the n -gram-likeness of LM surprisal

This analysis follows Michaelov and Levy (2026), who showed that Pythia checkpoints with stronger correlations to unigram, bigram, and trigram surprisal also tended to show stronger correlations with eye-tracking measures in the Provo Corpus (Luke and Christianson, 2018). In the present study, we ask whether the same qualitative pattern holds for Mandarin Chinese reading time. If eye-movement measures are especially sensitive to local distributional statistics, then LLM-RT correlations should be highest at the checkpoints where LLM surprisal is most strongly aligned with n -gram surprisal.

We use the FFD, GD, and TRT eye-tracking data from HKP, indexing progressively more cumulative aspects of eye-movement behavior during reading. Unigram, bigram, and trigram surprisal were computed using the released Chinese Wikipedia

⁷See Appendix B for the full model specifications.

Corpus	#Subj	#Units	#Char	Unit Len	#Align	No Merge	Corpus Finer	LM Finer	Crossing	C/L Ratio
GECO-CN	30	540	101,233	187.5	47,996	24,204 (50.4%)	7,176 (15.0%)	13,399 (27.9%)	3,217 (6.7%)	0.54
HKP	98	41	9,066	221.1	5,117	4,077 (79.7%)	681 (13.3%)	316 (6.2%)	43 (0.8%)	2.16
MECO	39	12	3,836	319.7	1,836	1,155 (62.9%)	121 (6.6%)	497 (27.1%)	63 (3.4%)	0.24

Table 3: Corpus statistics and tokenization alignment between corpus and LM tokenizer. #Units is the number of text units submitted to alignment (chunks of the GECO-CN novel, passage chunks of HKP, and whole paragraphs of MECO), not the number of linguistic sentences. #Align is the total number of aligned word spans; the four condition counts sum to #Align. C/L Ratio is the ratio of corpus-finer to LM-finer spans. For MECO, #Subj counts the subjects present in the analyzed data.

Corpus	Measure	14M	70M	160M	410M	1.4B
GECO-CN	TRT	1216	2219	2506	2610	2642
	FFD	194	458	579	661	676
	GD	990	1086	1206	1249	1322
HKP	TRT	365	488	508	396	396
	FFD	135	128	110	83	78
	GD	221	237	227	179	170
MECO	TRT	100	103	100	68	28
	FFD	42	37	40	24	14
	GD	99	82	81	54	28

Table 4: Mean Δllh of the linear regression models over late-training checkpoints ($\geq 300k$ steps; 16 checkpoints per cell), by corpus, measure, and model size. Bold marks the best-performing size per row. Higher values mean that adding surprisal yields a larger improvement in fit.

KenLM model from the edugp/kenlm repository.

⁸ Thus, unigram surprisal estimates the surprisal of the target token without conditioning on previous tokens, bigram surprisal conditions on the immediately preceding token, and trigram surprisal conditions on the two preceding tokens.

For each model checkpoint, we computed two sets of Pearson correlations: the correlation between LM surprisal and each reading-time measure; and the correlation between LM surprisal and each n -gram surprisal measure.

4 Results

4.1 Surprisal improves predictions in all three corpora

The results are presented in Figure 2, with corpora and reading time measures in different facets, and in Table 4, which summarizes late-training perfor-

⁸KenLM is an efficient toolkit for querying probabilistic n -gram language models (Heafield, 2011). For a target token w_i , n -gram surprisal was defined as negative log probability:

$$S_n(w_i) = -\log P(w_i \mid w_{i-n+1}, \dots, w_{i-1}).$$

mance. The first clear pattern we observed is that including surprisal in the regression model improves predictive power relative to the baseline model, as evidenced by the positive Δllh values throughout training, for all model sizes and in all three corpora. Unlike findings from a previous study that found no such effect on a small Chinese eye-tracking dataset (Xu et al., 2023), our result aligns with the general pattern found on cross-linguistic studies, especially the English ones (Oh and Schuler, 2023a,b; Wilcox et al., 2020, 2023b; Shain et al., 2024).

4.2 Scaling with model size and training is corpus-specific

We then observed strong corpus-specific effects in how predictive power scales with model size and training (Table 4). In GECO-CN, Δllh values increase sharply as model size and training data increase, and then reach a high plateau within about 1–4 billion training tokens; the largest model is the best predictor for every measure (TRT Δllh : 1216 for 14M vs. 2642 for 1.4B), and the top-end gain from 410M to 1.4B is small. Better fit to human reading times is thus obtained by the more powerful models – those with more parameters and trained on more data – up to a plateau. However, HKP shows the opposite pattern, exhibiting inverse scaling with respect to both training tokens and model size: for every model above 14M, predictive power peaks after about 0.2–2.5 billion training tokens and declines with further training, and late-training TRT Δllh peaks at 508 for 160M and falls to 396 for both 410M and 1.4B; for FFD and GD the two largest models are the weakest predictors of all. In MECO, the divergence between smaller and larger models is even more evident. Larger models reach the highest mid-training peaks (Figure 2), but their late-training predictive power collapses with size (TRT Δllh : 100 at 14M vs. 28 at 1.4B; GD: 99 vs. 28), inverting the size ordering at the largest sizes – a clear bigger-is-worse effect once model size becomes sufficiently large.

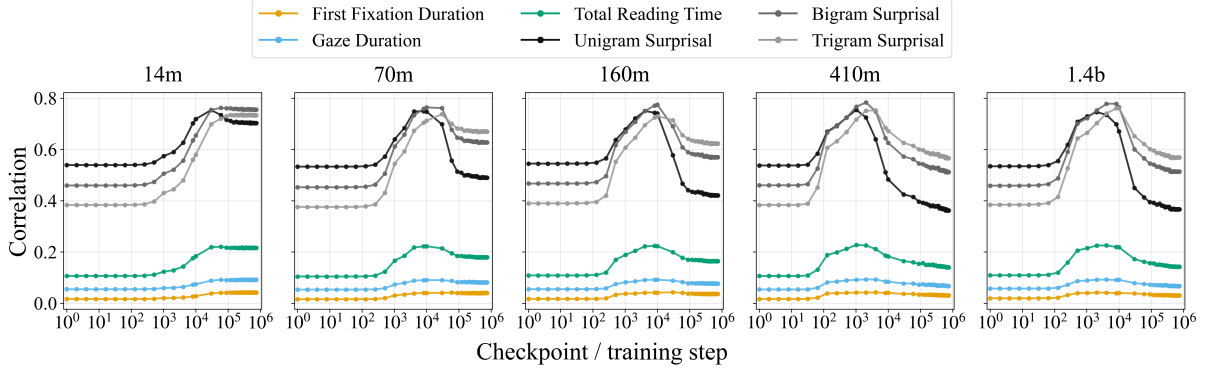


Figure 1: Training trajectories of LLM-RT and LLM- n -gram correlations in HKP. Each panel corresponds to a Chinese Pythia model size.

4.3 Early versus late reading-time measures

Across eye-tracking measures, FFD shows the weakest gains from surprisal, as reflected in its smaller Δllh values (Figure 2 and Table 4). This suggests that surprisal is less predictive of early fixation behavior than of later reading-time measures.

Besides, in HKP, scaling up harms early processing more than late processing: FFD Δllh falls by 42% from 14M to 1.4B (from 135 to 78), whereas TRT Δllh is slightly higher at 1.4B than at 14M (396 vs. 365). In MECO, by contrast, larger models lose predictive power on all three measures (FFD from 42 to 14, GD from 99 to 28, TRT from 100 to 28), and in GECO-CN, FFD continues to improve with model size alongside later measures.

4.4 Reading-time fit tracks the n -gram-likeness of surprisal

Overall, LM surprisal correlations with both reading times (LM-RT) and n -gram surprisal (LM- n -gram) rise early in training and decline after an initial peak (Figure 1). This post-peak degradation increases with model size: the LM-TRT correlation drops 0.04 below peak at 70M and 0.08 at 1.4B, compared to a simple plateau at 14M.

Across training checkpoints, LM-RT and LM- n -gram correlations move synchronously (Figure 3). Pooled over sizes, LM-RT correlations strongly track higher-order n -grams ($r \geq 0.85$ for bigrams and trigrams) but dissociate from unigrams ($r = 0.22$ – 0.56). Furthermore, early measures (FFD and GD) align primarily with trigram surprisal ($r = 0.99$ and 0.98), while TRT aligns with bigram surprisal ($r = 0.98$). Absolute LM-RT correlations remain low overall (maxima: 0.04 for FFD, 0.09 for GD, 0.23 for TRT), falling well short of values reported for the English Provo Corpus

(Michaelov and Levy, 2026).

The same pattern governs predictive power (Δllh): within each scale, checkpoints with higher trigram-likeness achieve superior Δllh . This coupling remains consistently high for early measures (FFD: $r = 0.98$; GD: $r = 0.91$ – 0.99), but attenuates with scale for TRT ($r = 0.98$ at 14M vs. 0.66 at 1.4B). Conversely, unigram-likeness predicts lower Δllh for models above 14M (TRT: $r = -0.43$ to -0.67). Finally, peak Δllh occurs much later in pretraining than peak raw correlation (e.g., step 60k vs. 4096 for 1.4B TRT).

5 General Discussion

Adding LM-derived surprisal to baseline mixed-effects regressions improved fit for every reading-time measure in all three corpora. Surprisal is thus predictive of Mandarin Chinese reading times, contrary to earlier null findings. How this predictive power scales with model size and training, however, differed across corpora: it grew with scale in GECO-CN and inverted in HKP and MECO. Below, we relate our results to earlier studies of Mandarin Chinese (Section 5.1), characterize the corpus-specific scaling (Section 5.2), weigh a candidate explanation for the post-peak decline (Section 5.3), and assess text genre and reading regime as moderators (Section 5.4).

5.1 Relation to earlier null findings

Earlier studies found no reliable effect of surprisal on Chinese reading times (Xu et al., 2023; Sun and Liu, 2025). Several differences could explain the discrepancy. Xu et al. (2023) estimated surprisal with GPT-2 and mGPT, neither trained primarily on Chinese, and analyzed a small dataset; our models were trained on Chinese text with a Chinese

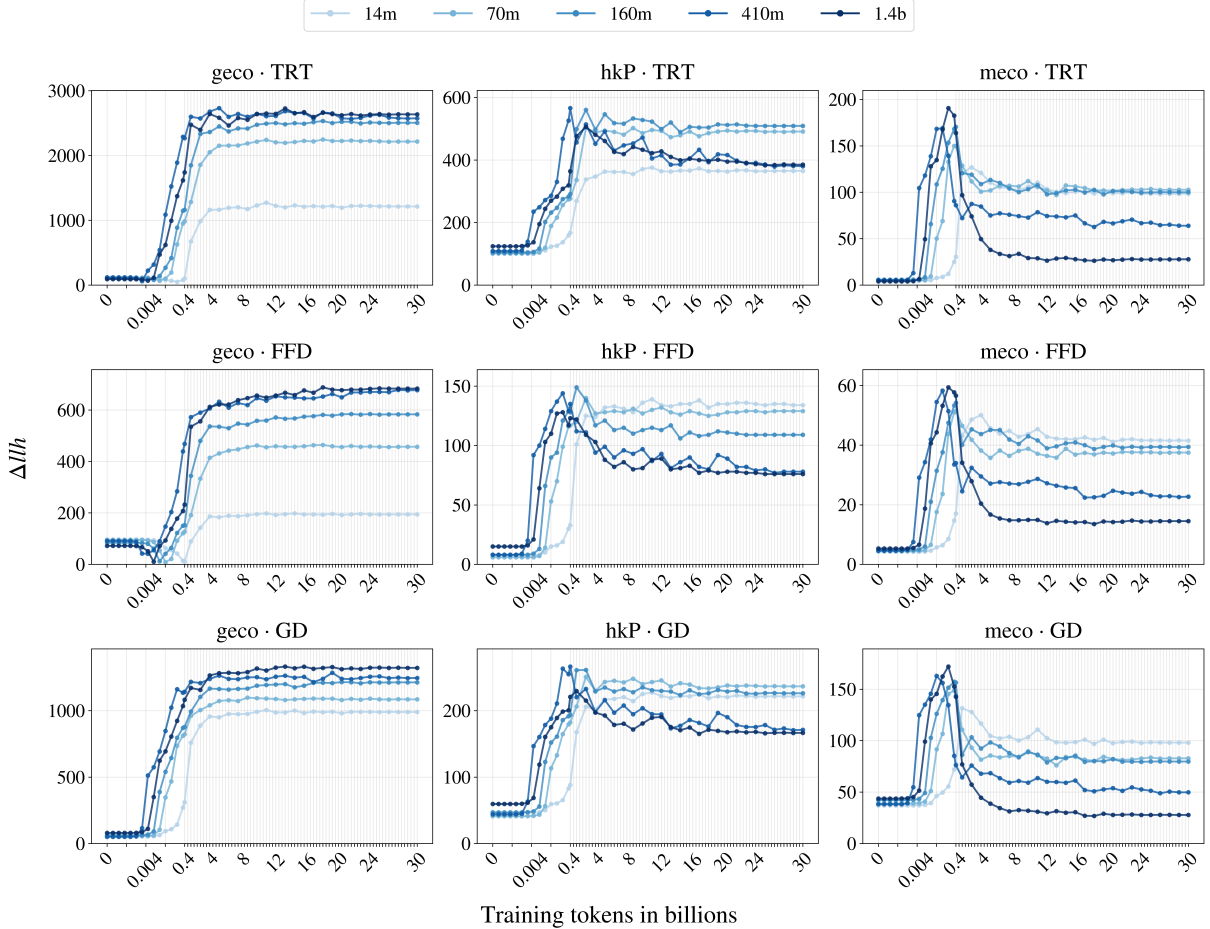


Figure 2: Δllh for reading times across training. Positive values indicate that including surprisal improves the prediction of reading times relative to the baseline.

tokenizer, our alignment scheme handles the frequent mismatches between corpus segmentation and LM tokenization, and we analyze three corpora that together contain more than 570,000 TRT observations (Table 1). The two sets of results may also be less contradictory than they appear. In HKP, raw correlations between LM surprisal and reading times are low (at most 0.04 for FFD, 0.09 for GD, and 0.23 for TRT), far below those reported for English (Michaelov and Levy, 2026). Since the effect is reliable but weak, it can be easy to miss with models not trained on Mandarin Chinese or with small samples. In addition, properties of reading Mandarin Chinese text may have contributed to this weakness. As explained in Section 3.4, Mandarin Chinese is unspaced and its word boundaries are not always clear-cut, so corpus-defined words may be less likely to coincide with the units whose predictability drives eye movements than in English or Dutch reading.

5.2 The direction of scaling depends on the corpus

The corpora largely agree about whether larger models can predict reading times better: in all three, fit rises steeply early in training and peaks or plateaus within the first few billion tokens, with the larger models generally reaching the higher values (Figure 2), although in HKP the 1.4B model peaks below the 160M and 410M models. Nevertheless, they disagree about the effect of further training. In GECO-CN, fit plateaus and the size ordering is preserved, with diminishing gains at the top end; in HKP, fit declines after the peak, most for the largest models, so that late-training fit peaks at small or intermediate sizes; and in MECO, the decline is proportionally steepest and inverts the size ordering at the largest sizes (Section 4). Because the Δllh curves have converged by the end of training, these patterns are asymptotic rather than artifacts of under-training. The timing of the peak also matches a finding for English: in HKP, the 1.4B model’s

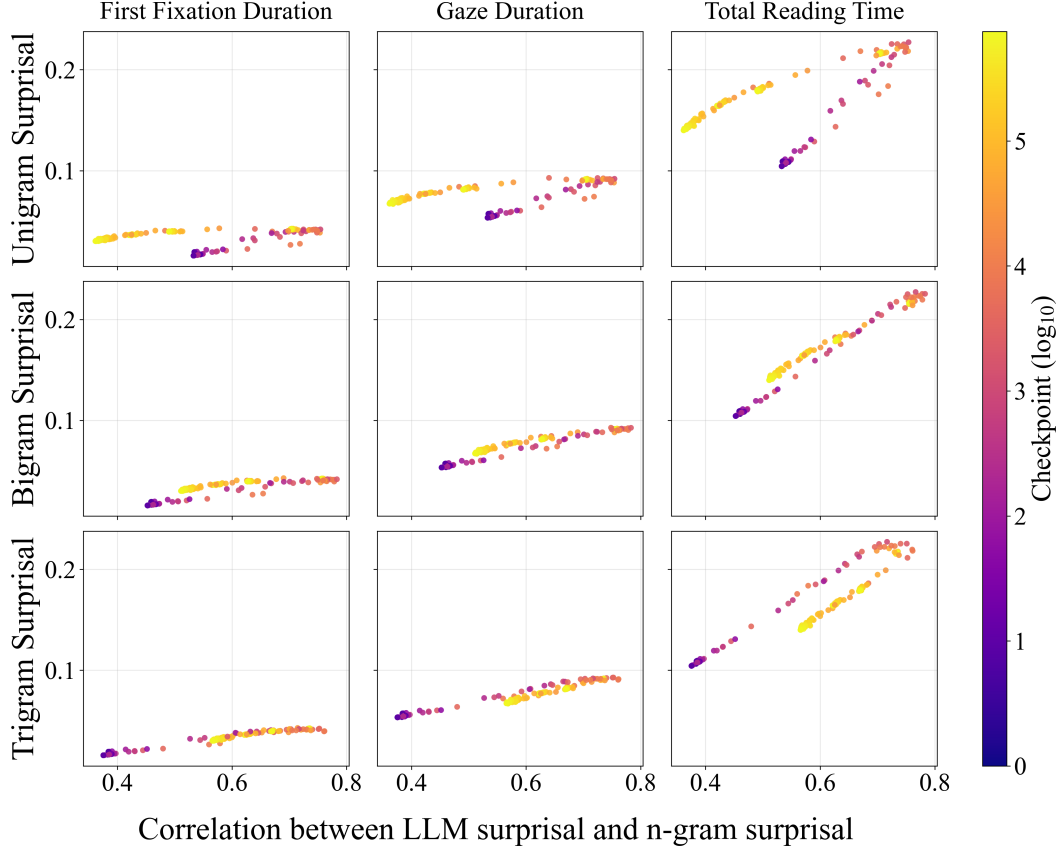


Figure 3: Correspondence between n -gram-likeness and reading-time fit in HKP. Each point is one model checkpoint. The x-axis shows the Pearson correlation between LM surprisal and n -gram surprisal; the y-axis shows the Pearson correlation between LM surprisal and a reading-time measure.

TRT fit peaks at step 60k, or about 2.5B training tokens, close to the roughly two billion tokens at which Transformer surprisal best predicts English reading times (Oh and Schuler, 2023a). Checkpoints at this stage also fit better than fully trained models in HKP (for all models above 14M) and in MECO, making them a safer default there; in GECO-CN, by contrast, fully trained models fit better at every size.

The measures differ as well. In HKP, the cost of scale falls mainly on early measures: from 14M to 1.4B, FFD Δllh drops by 42%, whereas TRT Δllh is slightly higher at 1.4B than at 14M (Table 4). This echoes De Varda and Marelli (2023), who found across several languages that larger models predicted early eye-tracking measures worse but late measures better. In MECO, however, all three measures fall by about two-thirds or more over the same range (FFD from 42 to 14, GD from 99 to 28, TRT from 100 to 28), so the asymmetry is not general. It also needs to be confirmed on matched spans: GD excludes multi-word spans

in all corpora and FFD excludes them in MECO, whereas TRT includes them, and in GECO-CN and HKP the FFD of a multi-word span is paired with surprisal that includes words not yet fixated.

These results bear on the scaling debate in English, where positive correlations between LM quality and predictive power for smaller models and training sets (Wilcox et al., 2020, 2023a) and inverse scaling for larger Transformers (Oh et al., 2022; Oh and Schuler, 2023b) invite an explanation in terms of model regime. Our models span the same 14M–1.4B range in every corpus, yet show both patterns, which also vary across reading-time measures. Part of the between-study disagreement may therefore reflect the materials and measures a study samples rather than the model regime alone.

5.3 A candidate explanation for the post-peak decline

The n -gram analysis of HKP (Section 4.4) offers one clue for the post-peak decline. Within each model size, checkpoints whose surprisal is more

trigram-like tend to predict reading times better, most consistently for FFD ($r = 0.98$; for TRT, r falls from 0.98 at 14M to 0.66 at 1.4B). This fits the proposal that models can become too good at next-word prediction (Michaelov and Levy, 2026): early training sharpens the local statistics to which eye movements are most sensitive (Oh and Schuler, 2025; Opedal et al., 2024), whereas later training may add higher-order regularities, such as knowledge of rare words and entities, that improve next-word prediction without making surprisal more human-like. Unigram-likeness, in contrast, relates negatively to TRT Δllh above 14M, as expected given that word frequency is already in the baseline model.

The analysis nonetheless supports a narrower claim than it may appear. Thirteen of the 40 checkpoints per model fall within the first 4,096 steps, where all measures of model quality rise together from a near-random start, so the checkpoint-level correlations largely reflect this shared early rise rather than the post-peak decline at issue. Indeed, for the 1.4B model, TRT Δllh peaks at step 60k, whereas the raw LM-TRT correlation, which moves with n -gram-likeness, peaks at step 4,096: fit keeps improving well after surprisal has begun to move away from n -gram statistics. In addition, the n -gram model was trained on Chinese Wikipedia rather than on our training data, so n -gram-likeness partly reflects similarity to Wikipedia text. Restricting the analysis to post-peak checkpoints with item-bootstrapped intervals, decomposing surprisal at the word level into trigram-predicted and residual components, and comparing checkpoints against the cloze norms of the Beijing Sentence Corpus (Pan et al., 2022) would test the account more directly. Most importantly, because the analysis covers only HKP, it cannot yet explain why scaling differs across corpora. It does make a falsifiable prediction: in GECO-CN, larger models should stay close to n -gram statistics late in training; if they instead drift away without losing fit, n -gram-likeness cannot be what distinguishes the corpora.

5.4 Text genre and reading regime

One possibility is that expository text encourages skimming, which would loosen the link between predictability and eye movements. MECO, the encyclopedic corpus, does not support this: it was read most slowly (mean TRT 359 ms vs. 282 ms in GECO-CN; Table 1), had the smallest share of TRT observations removed as skipped or missing

(about 45%, vs. 51% in GECO-CN and 60% in HKP; Table 5), and follows each text with comprehension questions. HKP, which consists of news reports, gives mixed evidence: it had the largest share of skipped or missing observations, but was still read more slowly than GECO-CN (mean TRT 290 vs. 282 ms). More fundamentally, a looser link between predictability and eye movements would predict weaker surprisal effects overall, not a different direction of scaling: surprisal from small models predicts reading times in all three corpora, and what differs is whether larger, longer-trained models improve on it. Genre itself remains a candidate: GECO-CN, the only narrative corpus, is also the only one in which late-training predictive power increases with model size, but with a single narrative corpus, genre is confounded with everything else specific to GECO-CN. Taken together, our data do not yet explain why scaling differs across corpora: the n -gram account has been tested only in HKP, a skimming account is not supported, genre cannot be separated from the other properties of GECO-CN, and other candidates, such as text difficulty, lexical composition, and overlap with the training data, remain untested.

6 Conclusion

This study asked whether surprisal estimated from decoder LMs predicts reading times in Chinese, how model scale modulates this predictive power, and how the corpus-specific scaling can be explained. We introduced an alignment scheme that resolves the mismatch between the word segmentation of eye-tracking corpora and the subword tokenization of LMs in Chinese, and we pretrained a Chinese-Pythia suite (14M–1.4B parameters, 30B tokens) with 40 checkpoints per size to trace predictive power across training. Across three paragraph-level eye-tracking corpora and three measures, adding surprisal consistently improved fit over baselines with word length and frequency: surprisal is predictive of Chinese reading times, contrary to earlier null findings, although its raw correlations with reading times in HKP are modest.

How predictive power scales, however, is corpus-specific. In all three corpora, fit peaks or plateaus within the first few billion training tokens; what differs is the effect of further training. In GECO-CN, fit plateaus, and late-training fit increases monotonically with model size; in HKP, fit declines after the peak, most for the largest models; and in MECO,

the largest models reach the highest mid-training peaks and then fall furthest, inverting the size ordering at the largest sizes. No single scaling law – bigger-better or bigger-worse – holds across Chinese corpora, even within a single model suite. A follow-up analysis on HKP found that checkpoints whose surprisal is more trigram-like tend to predict reading times better, most consistently for FFD, in line with the proposal that models can become too good at next-word prediction; however, this coupling largely reflects the shared rise early in training, and whether it explains the differences between corpora remains to be tested.

We therefore conclude that whether larger or better-trained models predict reading better depends on the materials, not on model scale alone. Candidate explanations include text difficulty, lexical composition, overlap with training data, and genre. For psycholinguistic modeling in Chinese and beyond, this cautions against generalizing scaling conclusions from a single corpus and motivates treating corpus characteristics as a first-class factor in model–human alignment research.

Limitations

This study has three main limitations. First, the n -gram-likeness analysis (Section 4.4) is restricted to HKP and supports only a correlational claim; extending it to GECO-CN and MECO is a natural next step. Second, Δllh magnitudes scale with the number of observations and are not directly comparable across corpora; all comparisons in this work are therefore made within a corpus, and simulation-based or normalized effect-size measures would be needed for genuine cross-corpus quantification. Third, genre may moderate how surprisal’s predictive power scales (Section 5.4), but the present design cannot isolate it: GECO-CN is the only narrative corpus in this study, so genre is confounded with everything else specific to it.

Ethical Considerations

This study uses eye-tracking data from previously published corpora and involves no new human-subject data collection. The Chinese-Pythia models were trained on publicly available text (CCI3-HQ) and are released for research purposes.

Acknowledgments

This project is supported by the National Social Science Fund of China awarded to Hai Hu

(25BYY133).

References

- Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O’Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International conference on machine learning*, pages 2397–2430. PMLR.
- Sam Boeve and Louisa Bogaerts. 2025. A systematic evaluation of dutch large language models’ surprisal estimates in sentence, paragraph and book reading. *Behavior Research Methods*, 57(9):266.
- Yiming Cui, Ziqing Yang, and Xin Yao. 2024. [Efficient and effective text encoding for chinese llama and alpaca](#). Preprint, arXiv:2304.08177.
- Andrea De Varda and Marco Marelli. 2023. Scaling in cognitive modelling: A multilingual approach to human reading times. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 139–149.
- Xi Fan and Ronan Reilly. 2020. Reading development at the text level: an investigation of surprisal and embedding-based text similarity effects on eyemovements in chinese early readers. *Journal of eye movement research*, 13(6):10–16910.
- John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the North American Chapter of the Association for Computational Linguistics*.
- Patrick Haller, Lena Bolliger, and Lena Jäger. 2024. [Language models emulate certain cognitive profiles: An investigation of how predictability measures interact with individual differences](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7878–7892, Bangkok, Thailand. Association for Computational Linguistics.
- Kenneth Heafield. 2011. [KenLM: Faster and smaller language model queries](#). In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.
- Tatsuki Kuribayashi, Yohei Oseki, Ana Brassard, and Kentaro Inui. 2022. [Context limitations make neural language models more human-like](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10421–10436, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, and Kentaro Inui. 2021.

- Lower perplexity is not always human-like. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5203–5217, Online. Association for Computational Linguistics.
- Tatsuki Kuribayashi, Yohei Oseki, Souhaib Ben Taieb, Kentaro Inui, and Timothy Baldwin. 2025. [Large language models are human-like internally](#). *Transactions of the Association for Computational Linguistics*, 13:1743–1766.
- Roger Levy. 2008. [Expectation-based syntactic comprehension](#). *Cognition*, 106(3):1126–1177.
- Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O’Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. [Few-shot Learning with Multilingual Language Models](#). *Preprint*, arXiv:2112.10668.
- Steven G. Luke and Kiel Christianson. 2018. [The Provo Corpus: A large eye-tracking corpus with predictability norms](#). *Behavior Research Methods*, 50(2):826–833.
- Wei-Yun Ma and Keh-Jiann Chen. 2003. [Introduction to CKIP chinese word segmentation system for the first international chinese word segmentation bakeoff](#). In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, pages 168–171.
- Wei-Yun Ma and Chu-Ren Huang. 2006. [Uniform and effective tagging of a heterogeneous giga-word corpus](#). In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC’06)*, Genoa, Italy. European Language Resources Association (ELRA).
- Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. [Revisiting the uniform information density hypothesis](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 963–980.
- James A. Michaelov and Roger P. Levy. 2026. [N-gram-like language models predict reading time best](#). *arXiv preprint arXiv:2603.09872*.
- Byung-Doh Oh, Christian Clark, and William Schuler. 2022. [Comparison of structural parsers and neural language models as surprisal estimators](#). *Frontiers in Artificial Intelligence*, 5.
- Byung-Doh Oh and William Schuler. 2023a. [Transformer-based language model surprisal predicts human reading times best with about two billion training tokens](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Findings)*, pages 1915–1921, Singapore. Association for Computational Linguistics.
- Byung-Doh Oh and William Schuler. 2023b. [Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times?](#) *Transactions of the Association for Computational Linguistics*, 11:336–350.
- Byung-Doh Oh and William Schuler. 2025. [Dissociable frequency effects attenuate as large language model surprisal predictors improve](#). *Journal of Memory and Language*, 143:104645.
- Byung-Doh Oh, Shisen Yue, and William Schuler. 2024. [Frequency explains the inverse correlation of large language models’ size, training data amount, and surprisal’s fit to reading times](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2644–2663, St. Julian’s, Malta. Association for Computational Linguistics.
- Byung-Doh Oh, Hongao Zhu, and William Schuler. 2025. [The inverse scaling effect of pre-trained language model surprisal is not due to data leakage](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1820–1827, Vienna, Austria. Association for Computational Linguistics.
- Andreas Opedal, Eleanor Chodroff, Ryan Cotterell, and Ethan Wilcox. 2024. [On the role of context in reading time prediction](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3042–3058. Association for Computational Linguistics.
- Jinger Pan, Ming Yan, Eike M Richter, Hua Shu, and Reinhold Kliegl. 2022. [The beijing sentence corpus: A chinese sentence corpus with eye movement data and predictability norms](#). *Behavior Research Methods*, 54(4):1989–2000.
- Tiago Pimentel, Clara Meister, Ethan G Wilcox, Roger P Levy, and Ryan Cotterell. 2023. [On the effect of anticipation on reading times](#). *Transactions of the Association for Computational Linguistics*, 11:1624–1642.
- Team Qwen, , An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Keith Rayner. 1998. [Eye movements in reading and information processing: 20 years of research](#). *Psychological Bulletin*, 124(3):372–422.

- Cory Shain, Idan A Blank, Evelina Fedorenko, Edward Gibson, and William Schuler. 2022. Robust effects of working memory demand during naturalistic language comprehension in language-selective cortex. *Journal of Neuroscience*, 42(39):7412–7430.
- Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. 2024. [Large-scale evidence for logarithmic effects of word predictability on reading time](#). *Proceedings of the National Academy of Sciences*, 121(10):e2307876121.
- Noam Siegelman, Sascha Schroeder, Cengiz Acartürk, Hee-Don Ahn, Svetlana Alexeeva, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, et al. 2022. Expanding horizons of cross-linguistic research on reading: The multilingual eye-movement corpus (meco). *Behavior research methods*, 54(6):2843–2863.
- Noam Siegelman, Sascha Schroeder, Yaqian Borogjoon Bao, Cengiz Acartürk, Niket Agrawal, Lena S. Bolliger, Jan Brasser, César Campos-Rojas, Denis Drieghe, Dušica Filipović Đurđević, Sofya Goldina, Romualdo Ibáñez Orellana, Lena A. Jäger, Ómar I. Jóhannesson, Anurag Khare, Nik Kharlamov, Hanne B. S. Knudsen, Árni Kristjánsson, Charlotte E. Lee, Jun Ren Lee, Marina P. T. Leite, Simona Mancini, Nataša Mihajlović, Ksenija Mišić, Miloslava Orekhova, Olga Parshina, Milica Popović Stijačić, Athanassios Protopapas, David R. Reich, Anurag Rimzhim, Rui Rothe-Neves, Thais M. M. Sá, Andrea Santana-Covarrubias, Irina Sekerina, Heida M. Sigurdardottir, Anna Smirnova, Priyanka Srivastava, Elisangela N. Teixeira, Ivana Ugrinic, Kerem Alp Usal, Karolina Vakulya, Ark Verma, João M. M. Vieira, Denise H. Wu, Jin Xue, Sunčica Zdravković, Junjing Zhuo, Laoura Ziaka, and Victor Kuperman. 2025. [Wave 2 of the multilingual eye-movement corpus \(MECO\): New text reading data across languages](#). 12(1):1183.
- Nathaniel J Smith and Roger Levy. 2013. [The effect of word predictability on reading time is logarithmic](#). *Cognition*, 128(3):302–319.
- Longjiao Sui, Nicolas Dirix, Evy Woumans, and Wouter Duyck. 2023. [GECO-CN : Ghent Eye-tracking Corpus of sentence reading for Chinese-English bilinguals](#). *BEHAVIOR RESEARCH METHODS*, 55(6):2743–2763.
- Kun Sun and Haitao Liu. 2025. Attention-aware semantic relevance predicting chinese sentence reading. *Cognition*, 255:105991.
- Liangdong Wang, Bo-Wen Zhang, Chengwei Wu, Hanyu Zhao, Xiaofeng Shi, Shuhao Gu, Jijie Li, Quanyue Ma, Tengfei Pan, and Guang Liu. 2024. Cci3.0-hq: a large-scale chinese dataset of high quality designed for pre-training large language models. *arXiv preprint arXiv:2410.18505*.
- Ethan G. Wilcox, Clara Meister, Ryan Cotterell, and Tiago Pimentel. 2023a. [Language model quality correlates with psychometric predictive power in multiple languages](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7503–7511. Association for Computational Linguistics.
- Ethan G Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P Levy. 2023b. [Testing the predictions of surprisal theory in 11 languages](#). *Transactions of the Association for Computational Linguistics*, 11:1451–1470.
- Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the 42th Annual Meeting of the Cognitive Science Society - Developing a Mind: Learning in Humans, Animals, and Machines, CogSci 2020*.
- Ethan Gotlieb Wilcox, Michael Y Hu, Aaron Mueller, Alex Warstadt, Leshem Choshen, Chengxu Zhuang, Adina Williams, Ryan Cotterell, and Tal Linzen. 2025. Bigger is not always better: The importance of human-scale language modeling for psycholinguistics. *Journal of Memory and Language*, 144:104650.
- Simon N Wood. 2025. Generalized additive models. *Annual Review of Statistics and Its Application*, 12(1):497–526.
- Yushu Wu and Chunyu Kit. 2023. [Hong kong corpus of chinese sentence and passage reading](#). *Scientific Data*, 10(1):899.
- Weijie Xu, Jason Chon, Tianran Liu, and Richard Futrell. 2023. The linearity of the effect of surprisal on reading times across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15711–15721.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.
- Guangyao Zhang, Panpan Yao, Guojie Ma, Jingwen Wang, Junyi Zhou, Linjieqiong Huang, Pingping Xu, Lijing Chen, Songlin Chen, Junjuan Gu, et al. 2022a. The database of eye-movement measures on words in chinese reading. *Scientific Data*, 9(1):411.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022b. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

A Details on the three eye-tracking datasets

- Ghent Eye-tracking Corpus for Chinese-English bilinguals, **GECO-CN** (Sui et al., 2023) contains eye-movement data from 30 native Chinese (L1) speakers reading the detective novel *The Mysterious Affair at Styles* by Agatha Christie (1920). Translation was revised by the first author, who is a native Chinese speaker. The published Chinese version of the novel comprises 59,403 words and 5,053 unique types; our extraction of GECO-CN comprises 895,938 observations from the 30 subjects.
- Hong Kong Corpus of Chinese sentence and passage reading, **HKC** (Wu and Kit, 2023) releases eye-tracking data of 98 L1 speakers of Chinese, who read 300 single-line sentences and 7 multiline passages selected from the XIN subcorpus of the Chinese Gigaword Corpus (Ma and Chen, 2003; Ma and Huang, 2006), with no overlap in between. The single-line sentences contain 5,150 word tokens (1,354 types), while the passages include 4,967 word tokens (1,434 types). Taken together, HKC comprises 10,117 word tokens (2,160 types), covering a variety of topics including celebrity news, city development, education, etc. In our extraction, the seven passages (HKP, the subset analyzed in this work) comprise 41 sentence lines and 5,911 word tokens (1,582 types).
- Multilingual Eye-Movement Corpus, **MECO** (Siegelman et al., 2025) comprises eye-tracking data from native Chinese speakers reading 12 Wikipedia-like paragraphs; the published Chinese sample includes 77 participants reading texts in both simplified and traditional Chinese. We analyze the simplified-Chinese texts, for which our extraction contains 39 subjects, 71 sentences, 2,024 word tokens (1,051 types), and 55,901 observations.

Preprocessing proceeds in three steps, applied to each reading-time measure: (1) we exclude skipped

Corpus	Measure	Merged	Step 1	Step 2	Step 3
GECO-CN	TRT	719,518	353,763	353,328	348,105
	FFD	719,518	353,763	353,761	348,527
	GD	719,518	257,369	257,347	253,298
HKP	TRT	489,847	197,317	197,141	195,840
	FFD	489,847	197,355	197,352	196,048
	GD	489,847	159,373	159,360	158,274
MECO	TRT	49,730	27,521	27,420	27,192
	FFD	49,730	26,527	26,527	26,293
	GD	49,730	26,517	26,517	26,283

Table 5: Number of observations after each step of pre-processing, starting from the merged word-level dataset (after tokenization alignment; Section 3.4). Step 1 excludes skipped words (reading time of 0 or missing); step 2 excludes words with non-finite surprisal or reading times of 2000 ms or longer; step 3 excludes paragraph-initial words. GD counts drop sharply at step 1 because GD is undefined for multi-corpus-token spans. Counts are identical across all model sizes and checkpoints.

words, i.e., words with a reading time of 0 or missing; (2) we exclude words with non-finite surprisal or reading times of 2000 ms or longer; (3) we exclude paragraph-initial words. We do not apply a hard surprisal cutoff, as it would remove multi-token words in Chinese disproportionately; words attached to punctuation and sentence-final words are retained.

The number of observations left after each step is presented in Table 5.

B Regression models used in this study

Details of the regression models reported in the main section are presented below.⁹ RT is the reading time (i.e., FFD, GD, TRT) of the word of interest (w_i). *surp.s* refers to the standardized surprisal of w_i , while *wlen.s* and *logfreq.s* respectively denote the standardized word length and standardized log-transformed word frequency of w_i ; a suffix 1 (e.g., *surp1.s*, *wlen1.s*) denotes the corresponding spillover term for the preceding word. Δllh refers to the change in log-likelihood, calculated as the difference between regression models with and without surprisal.

Predictive power of surprisal (Section 4) For the tests assessing the effect of surprisal within each corpus, we use the following model:

$$\text{lmer}(\text{RT} \sim \text{surp.s} + \text{surp1.s} + \text{wlen.s} + \text{wlen1.s} + \text{logfreq.s} + \text{logfreq1.s} + (\text{wlen.s} | \text{subject}))$$

⁹We modify the code base of Xu et al. (2023) to run the regression analyses in this study.

The baseline model is identical except that the surprisal terms are removed:

```
lmer(RT ~ wlen.s + wlen1.s +  
      logfreq.s + logfreq1.s +  
      (wlen.s|subject))
```

Both models are fit with maximum likelihood.