Modeling Bottom-up Information Quality during Language Processing

Anonymous ACL submission

Abstract

Contemporary theories model language processing as integrating both top-down expectations and bottom-up inputs. One major prediction of such models is that the quality of the bottom-up inputs modulates ease of processing—noisy inputs should lead to difficult and effortful comprehension. We test this prediction in the domain of reading. First, we propose an information-theoretic operationalization for the "quality" of bottom-up information as the mutual information (MI) between visual information and word identity. We formalize this prediction in a mathematical model of reading as Bayesian update. Second, we test our operationalization by comparing participants' reading times in conditions where words' information quality has been reduced, either by occluding their top or bottom half, with full words. We collect data in English and Chinese. We then use multimodal language models to estimate the mutual information between visual inputs and words. We use these data to estimate the specific effect of reduced information quality on reading times. Finally, we compare how information is distributed across visual forms. In English and Chinese, the upper half contains more information about word identity than the lower half. However, the asymmetry is more pronounced in English, a pattern which is reflected in the reading times.

1 Introduction

During reading, individuals actively expend cognitive effort to extract information. Many contemporary theories of language comprehension in general, and reading in particular, model this process as a rational integration of bottom-up and top-down information (Legge et al., 1997; Norris, 2006; Bicknell and Levy, 2010; Gibson et al., 2013; Gauthier and Levy, 2023). Bottom-up information refers to the perceptual input (e.g., visual forms of words), while top-down information includes the prior beliefs and expectations about what messages or word-forms are likely to be encountered, and is guided by the reader's linguistic and contextual knowledge. A central prediction of such models is that the ease of reading should be influenced by the quality of the bottom-up information. In the modality of visual reading, visual signals that effectively convey information about the intended message are expected to facilitate fast and effortless comprehension. Conversely, degraded visual signals—caused by factors such as lighting, occlusion, or visual interference—are likely to increase processing effort and raise the likelihood of errorful reading.

This prediction fits well within noisy channel models of reading. In a noisy-channel model (Shannon, 1948), a message is encoded and sent over a channel, where it is potentially corrupted. A receiver, at the other end of the channel, must decode the most probable intended message given the received inputs. Previous work has looked at the role of noise during reading, demonstrating how noise over uncertain inputs can lead to non-veridical interpretations (Levy, 2008b; Gibson et al., 2013).

While intuitive, to the best of our knowledge, this prediction has not been quantified within a formal computational model of reading. That is, although many theories of reading assume that poorer sensory input leads to more effortful processing, they have not derived or test this relationship quantitatively. In this paper, we aim to fill this gap by providing an informationtheoretically grounded, quantitative account of how bottom-up input quality affects processing effort. Our central proposal is that input quality can be formalized as the mutual information (MI) between (visual) input and word identity. From an information-theoretic perspective, a signal is informative to the extent that it reduces uncertainty about a target variable—in this case, the identity of a word. We assume that greater effort manifests in longer reading times, and therefore predict that reductions in mutual information should lead to systematic slowdowns in reading.

This paper makes three contributions: First, we instantiate the above operationalization of visual input quality in reading under a formal model of reading as a Bayesian update. Second, we provide a quantitative estimate of the cost of reduced input quality on processing effort. To do so, we use multimodal language models to estimate mutual information over a dataset of partially masked word images. We then collect human reading times on the same stimuli, using the MoTR paradigm (Wilcox et al., 2024), which simulates eye-tracking, and can be used to collect data over the web. We use these data to estimate the relationship as a specific slowdown in terms of nats of mutual information per millisecond of processing time. Our data suggest that the cost of reduced information is not linear—small losses in MI can lead to disproportionately large increases in reading time, particularly in the upper ranges of a signal's informational range.

Our third contribution is to compare how information is distributed across visual forms of words in two typologically distinct languages. To that end, we collect data in both English and Chinese. We find that, in both languages, the upper half of a word contains more information about word identity than the lower half. However, the asymmetry is more pronounced in English than in Chinese, a pattern that is reflected in the reading times.

2 Formal Model

2.1 Reading as Bayesian Update

Following an extensive prior literature (Norris, 2006; Bicknell and Levy, 2010; Gauthier and Levy, 2023), we model word recognition as a Bayesian update process. We model comprehension as being over words drawn from a vocabulary $w \in \mathcal{W}$, where W is a variable that ranges over words. We refer to a word at a particular timestep, t as w_t and the random variable ranging over words at this timestep as W_t. We assume that readers intake individual samples of input $e \in \mathbb{R}$, where E is a variable ranging over samples[1]. These can be either a patch of visual input for visual reading or a haptic percept in the case of braille. Following previous work (Bicknell and Levy, 2010), we model the process of reading as one of sequential word identification given input e and a previous context of words $\mathbf{w}_{<t}$. In such models, readers are assumed to rationally integrate their prior expectations about a word, $P(w_t \mid \mathbf{w}_{<t})$, with the likelihood of the observed input, $P(e_i \mid w_t, \mathbf{w}_{<t})$. Instead of a single sample, we assume that readers integrate evidence over k samples. The rational update process we use to model reading is therefore:

$$P(w_t \mid \mathbf{e}_{1:k}, \mathbf{w}_{<t}) \propto \qquad (1)$$

$$P(w_t \mid \mathbf{w}_{<t}) \times \prod_{i=1}^{k} P(e_i \mid w_t, \mathbf{w}_{<t})$$

This tells us how readers update beliefs about a word given inputs and priors. But reading is a dynamic process. How do readers choose when to move on to the next word? We propose that readers draw samples until the uncertainty about the current word reaches a threshold, ϕ, at which point they move on. We quantify uncertainty as the entropy of the posterior distribution. That is, sampling continues until:

$$H(P(w_t \mid \mathbf{e}_{1:k}, \mathbf{w}_{<t})) \leq \phi \qquad (2)$$

However, given a particular actual input w^* we cannot be certain how many samples a reader draws or what information each sample contains. Therefore, for a given piece of text, we predict readers to move on when the *expected* entropy falls below this threshold, where the expectation is taken over uncertain inputs:

$$\mathbb{E}_{\mathbf{E}_{1:k}}[H(W_t \mid \mathbf{E}_{1:k}, \mathbf{w}_{<t})] \leq \phi \qquad (3)$$

[1] For simplicity, we model inputs as continuous and univariate. However, we acknowledge that inputs may be more aptly modeled as multivariate and see this as an easy extension of the formal presentation given here.

where k now represents the expected number of samples. Although we assume that reading does take place given a context, for the rest of this section, we will drop the word-context term, $\mathbf{w}_{<t}$. We note that it would be easy to add this term back into the subsequent equations as a conditioning variable without changing the overall model.

2.2 Quality of Bottom-Up Evidence

Here, we are primarily interested in how the quality of the inputs impacts the reading process. We model the quality of the inputs as the mutual information between the inputs and the word identities, i.e., as $I(W;E)$. That is, high-quality inputs do a better job of reducing uncertainty over words. For a given word-identification step, we can write the mutual information between a word and the total number of samples drawn as $I(W;\mathbf{E}_{1:k})$. Using the chain rule of mutual information (Cover, 1999) and assuming that the samples $\mathbf{E}$ are drawn i.i.d. and, furthermore, that there is *conditional independence* between samples, given W, we can make the following simplifications:[2]

$$I(W;\mathbf{E}_{1:k}) = \sum_{i=1}^{k} I(W;E_i \mid \mathbf{E}_{1:i-1}) \quad (4a)$$

$$\text{i.i.d. samples} = \sum_{i=1}^{k} I(W;E_i) \quad (4b)$$

$$= k \times I(W;E) \quad (4c)$$

How is the mutual information between inputs and words related to the reading process, as described above? We assume that taking samples and processing these samples takes cognitive effort. Following previous work, we also assume a link between effort and time (Levy, 2008a; Hale, 2001). Therefore, the more samples, k, a reader needs to take in order to reduce uncertainty, the longer it will take them to read a given word.

We can now link the quality of inputs to our reading process through the definition of mutual information:

$$I(W;\mathbf{E}_{1:k}) = H(W) - H(W \mid \mathbf{E}_{1:k}) \quad (5)$$

[2]For more discussion of these assumptions, see Appendix A.

Plugging in the equality from 4c, and the definition of conditional entropy,[3] we rearrange the terms to get:

$$\mathbb{E}_{\mathbf{E}_{1:k}}[H(W \mid \mathbf{E}_{1:k})] = H(W) - k \times I(W;E) \quad (6)$$

That is, the expected entropy of the posterior distribution, given uncertain inputs, is a function of the entropy over words, the number of samples taken, and the mutual information between the samples and the words.

For our model of reading, we are interested in when the entropy of the posterior distribution is approximately ϕ. In particular, we are interested in how many samples must be drawn to reach this threshold, as this determines the effort (and therefore the time) required to reduce uncertainty enough to move on to the subsequent word. Substituting in our threshold parameter in and rearranging the terms, we have:

$$k \approx \frac{H(W) - \phi}{I(W;E)} \quad (7)$$

The number of samples required to reach the threshold grows with the entropy of the distribution over W. Likewise, it decreases with the mutual information between W and E. Because we assume a link between the number of samples, effort and time, this leads us to the following two predictions:

Prediction 1 ***Top-Down Processing & Entropy:*** *As the entropy of a word-position W increases, average reading time increases.*

Prediction 2 ***Bottom-up Processing & Mutual Information:*** *As the mutual information between words W and their visual representations E decreases, average reading time increases.*

In fact, Prediction 1 has already been investigated by Pimentel et al. (2023), whose results confirm our prediction. Pimentel et al. refer to the entropy over the next word, given a set of previous words $H(W_t \mid \mathbf{w}_{<t})$ as a word's *contextual entropy*. They find that as word-level contextual entropy increases, so too does reading time. For the rest of this paper, therefore, we are interested in testing Prediction 2, namely

[3]That is: $H(X \mid Y) = \mathbb{E}_Y[H(X \mid Y)]$.

Figure 1: Example showing a screen from a MoTR trial with our three different reading conditions.

whether the quality of bottom-up evidence, modeled as mutual information between words and visual information, affects word-by-word reading times. We outline our methods to do so in the following section.

3 Methods

3.1 Materials

We use a portion of the OneStopQA dataset (Berzak et al., 2020). This dataset contains Guardian news articles, along with high-quality reading comprehension questions, which are linked to individual spans in the text. We selected three articles for inclusion in our study. One member of our research team with previous experience in English-Chinese translation hand-translated these texts and their associated questions into Mandarin. This translated corpus, which we term the **Chinese OneStopQA**, will be released along with the publication of this article.

Creating Noisy Words To create noised reading conditions, we occluded (i.e., masked with white) either the upper or lower half of every word in the dataset. There are potentially many ways to noise text. Other options were occluding the first half or second half of words, as well as Gaussian noise. Previously, Pimentel et al. (2021) found that the beginnings of words carry more information than their end. However, we were worried that entirely removing some letters or characters would make reading too difficult or frustrating for our participants, and that the removal of letters or characters demands very careful handling. Removing upper or lower half retains some information about each character. In addition, unlike simply adding Gaussian noise, upper and lower half occlusion allowed us to investigate *where* information was localized in English and Chinese orthographic systems. Our strategy lead to two additional research questions:

Sub Research Question 1 *Is information split up differentially between the upper and lower halves of orthographic words?*

Sub Research Question 2 *Does the location of information in upper vs. lower half of orthographic words differ between languages?*

3.2 Data Collection

Mouse Tracking for Reading (MoTR) To test our main predictions, we need a way of measuring (average) human reading times in our different conditions. To do so, we use Mouse Tracking for Reading (MoTR; Wilcox et al., 2024). In a MoTR trial, a blurred text is presented on a screen. A small region around the tip of a user's mouse brings the text into focus. Participants move the mouse in order to incrementally reveal and read the text. Participant mouse location is recorded and used as a proxy for gaze location. The time-stamped x/y coordinates are then turned into incremental word-by-word reading times, similar to reading times in an eye-tracking while reading experiment. As with eye-tracking, there are several ways to compute reading times. For our main analysis, we use *gaze duration*, which is the total amount of time a user spent revealing a word during their first pass. Wilcox et al. (2024) show that MoTR reading times are strongly correlated with eye-tracking and self-paced-reading times MoTR has been used to collect data in English and Russian (Oğuz et al., 2025), but not in Chinese.

Participants We recruited 54 English and 57 Chinese speakers on Prolific, requiring a minimum approval rate of 98% and the corresponding language to be their first and native language. Participants were compensated 3.75 GBP for a median reading time of 25 minutes.

Procedure Each participant read the article paragraphs presented screen by screen, with each screen randomly assigned to one of three conditions: upper-half occluded (i.e., lower-half visible), lower-half occluded (i.e., upper-half visible), or unoccluded (see Figure 1). In addition to reading texts and answering comprehension questions, we asked participants to rate the ease of reading after finishing all the trials.

3.3 Mutual Information Estimation

In Section 2, our model concerns words, W, and (visual) evidence sampled by the reader, $\mathbf{E}$. However, we do not have direct access to this evidence. Instead, as a proxy for our visual evidence, we estimate the mutual information between words W and their orthographic representations, representation $\mathbf{o} \in \mathbb{R}^d$, where the random variable $\mathbf{O}$ ranges over representations of different words. Following Pimentel et al. (2020), we decompose the mutual information as

$$I(W;\mathbf{O}) = H(W) - H(W \mid \mathbf{O}) \tag{8a}$$
$$\approx H_\theta(W) - H_\theta(W \mid \mathbf{O}) \tag{8b}$$

and separately estimate each term.

We estimate **unconditional entropy** $H_\theta(W)$ with a maximum likelihood estimation of the unigram distribution of Chinese characters and English words. We take the $9,933$ unique Chinese characters included in the modern Chinese character database [4], and the $60,384$ English words in the SUBTLEXus database (Brysbaert and New, 2024), and look up their frequencies using the Python library *wordfreq* (Speer, 2022) that supports both languages and aggregates data from multiple large-scale corpora, including subtitles, Wikipedia, news, fiction, and web content. Normalizing the frequencies, we obtain the empirical distribution $p_\theta(w)$ and from it we can directly compute the entropy $H_\theta(W)$. The empirical entropies are 5.59 and 7.12 nats for Chinese characters and English words.

We estimate the **conditional entropy** $H_\theta(W \mid \mathbf{O})$ in two stages. First, we compute the word-entropy conditioned on a specific orthographic representation, $H_\theta(W;\mathbf{O} = \mathbf{o})$ for every word in our vocabulary. We refer to this as the **pointwise conditional entropy**. We compute this value by taking the expectation of the information content, or **surprisal** of the word given its orthographic representation $\iota_\theta(w \mid \mathbf{o})$, where $\iota_\theta(\cdot) = -\log p_\theta(\cdot)$. Given a model with parameters θ that can produce our probability distribution of interest, that is, $p_\theta(w \mid \mathbf{o})$, the pointwise conditional entropy is calculated as:

$$H_\theta(W \mid \mathbf{o}) \approx \sum_{w \in \mathcal{W}} p_\theta(w \mid \mathbf{o}) \iota_\theta(w \mid \mathbf{o}) \tag{9}$$

We then estimate conditional entropy as the expectation of the pointwise conditional entropy with respect to $\mathbf{O}$, following the identity $H(W \mid \mathbf{O}) = \mathbb{E}_{\mathbf{O}}[H(W \mid \mathbf{O} = \mathbf{o})]$. We take the expectation over a set of held-out test samples:

$$H_\theta(W \mid \mathbf{O}) \approx \frac{1}{N} \sum_{n=1}^{N} H_\theta(W \mid \mathbf{o}^n) \tag{10}$$

where $\mathbf{o}^n$ is the n^{th} orthographic representation in the test set.

We note that using these methods, we can estimate not only the mutual information $I(W;\mathbf{O})$, but also its half-pointwise variant, also called the **information gain (IG)**, for a particular orthographic representation, where $IG(W;\mathbf{o}) = H(W) - H(W \mid \mathbf{o})$. While our formal prediction is made in terms of mutual information, in Section 4.3, we use IG to investigate the relationship between information contained in individual visual inputs and their respective reading times.

In recent work, similar methods have been used to study the relationship between words (as represented by text) and prosody, or the melody of speech (Wolf et al., 2023; Regev et al., 2025; Wilcox et al., 2025). However, these previous works learn distributions over real-valued variables that represent pitch. We wish to learn distributions over discrete w-valued variables $p_\theta(w \mid \mathbf{o})$. To obtain this distribution, we use multimodal language models, which we fine-tune to produce conditionalized distributions over words, given visual inputs. We do so with the following methods:

Fine-Tuning Data We adapt the Python library *TRDG* [5] to generate images of Chinese

[4] https://lingua.mtsu.edu/chinese-computing/

[5] https://github.com/Belval/TextRecognitionDataGenerator

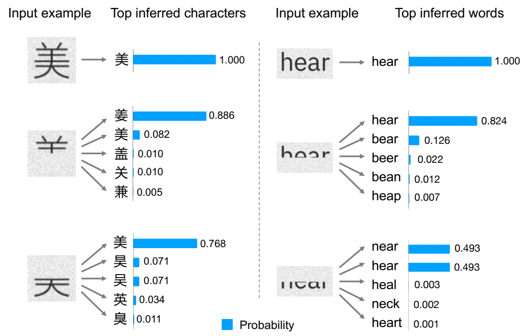

Figure 2: Results of our fine-tuned Qwen2.5 model for the Chinese character 美 ("beautiful") and the English word *hear*.

characters and English words from text, applying upper-, lower-half occlusion to create our different experimental conditions. We randomized font selection to enhance visual variability and added a small amount of Gaussian noise to the image backgrounds (Li et al., 2025). We generated $16,800$ Chinese character images and $44,800$ English word images for each of the three occlusion conditions as tuning data.

Predictive Multimodal Models We use three different model settings: First, we evaluate the pre-trained multimodal model Qwen2.5-VL-7B-Instruct[6] in a zero-shot setting. Qwen2.5-VL-7B is an open-source vision-language model developed by Alibaba, designed for high-accuracy multimodal analysis with enhanced visual understanding and text-image alignment (Wang et al., 2024; Bai et al., 2025). As top- and bottom-half occluded words are likely out-of-distribution with respect to the model's training data, we do not expect the mutual information estimate to be tight in this setting. For a better estimate, we then fine-tune Qwen2.5-VL-7B on our task-specific data to improve its performance. To complement the estimate from the pre-trained model, we also train a separate transformer-based OCR model (TransOCR) (Yu et al., 2023), from scratch, to perform the same prediction task. The model combines a ResNet encoder with a Transformer decoder for character recognition. Full training configurations and prompt designs for the Qwen and TransOCR models are provided in Appendix B and Appendix C, respectively.

To give a visual sense of how our models perform, Figure 2 shows sample images in our three experimental conditions, along with the predictions of the fine-tuned Qwen2.5 model.

4 Results

4.1 Human Reading Results

We show human reading times in Figure 3(a). In both languages, reading full words resulted in the shortest average reading times, as predicted. Interestingly, both languages follow a *Full* < *Upper* < *Lower* pattern, with lower-half visibility leading to the longest times. To quantify these effects, we fit linear mixed-effects models with visibility condition as a fixed effect, using sliding contrasts to compare *Upper* vs. *Full* and *Lower* vs. *Upper*. Random intercepts were included for subjects and items. In Chinese, both contrasts were significant: $\beta = 36.45\,\text{ms}$ and $\beta = 16.28\,\text{ms}$. In English, the effects were larger: $\beta = 54.64\,\text{ms}$ and $\beta = 90.06\,\text{ms}$[7]. All effects were statistically significant at $p < 0.001$.

These results can be interpreted as implying a visual asymmetry in both languages between ease of processing with respect to just upper and lower halves of words. The asymmetry is stronger in English, where the lower half leads to greater slowdowns. Participants' subjective ratings confirm this asymmetric pattern and further show that English lower halves are perceived as harder to read than Chinese ones (Appendix D).

4.2 Mutual Information Results

Figure 3(b) shows the information gain (IG) between word identity and visual input across the three visibility conditions, estimated by Qwen2.5-VL-7B-Instruct (zero-shot and fine-tuned) and TransOCR. Visually, these results show a decreasing IG trend among the *Full*, *Upper*, and *Lower* conditions. To test this statistically, we fit linear mixed-effects models for

[6] https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct

[7] Gaze duration was calculated for Chinese *characters* and English *words*, which may account for the generally longer reading times in English.

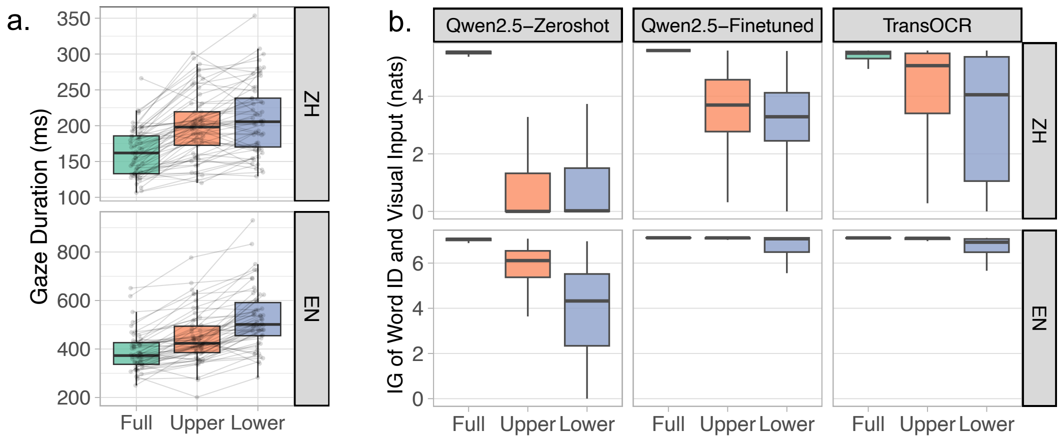

Figure 3: **(a)** Mean gaze durations measured in human reading under three visibility conditions. Boxes represent the interquartile range (middle 50%), center lines indicate the median, and whiskers show the overall data spread. Grey lines trace each participant's mean across conditions. EN: English; ZH: Simplified Chinese **(b)** Information gain (IG) between word identity and visual form under the three conditions, obtained with Qwen2.5 and TransOCR models.

each language–model pair, with visibility condition as a fixed effect and including a random intercept for item. As in the human reading analysis, we used sliding contrasts to compare our three conditions.

In Chinese, all models showed significant IG reductions when only the upper half was visible (Qwen2.5-Zeroshot: $\beta = -4.55$; Qwen2.5-Finetuned: $\beta = -1.85$; TransOCR: $\beta = -0.99$ nats), and IG from fine-tuned models dropped further when viewing changed from Upper to Lower (Qwen2.5-Finetuned: $\beta = -0.37$; TransOCR: $\beta = -1.01$ nats). In English, the zero-shot model showed the largest overall drop (Upper vs. Full: $\beta = -1.46$; Lower vs. Upper: $\beta = -2.11$ nats), while fine-tuned models showed smaller but consistent reductions (Qwen2.5-Finetuned: $\beta = -0.12, -0.47$; TransOCR: $\beta = -0.08, -0.35$ nats). All effects were statistically significant at $p < 0.001$. Panels (a) and (b) of Figure 3, taken together, reveal a clear pattern: as visual input degrades from *Full* to *Upper* to *Lower*, as measured by IG, reading times increase as well. We also obtained the mutual information $I(W; \mathbf{O})$ from the Qwen2.5 and TransOCR models, although we did not use them in our analysis above. The mutual information $I(W; \mathbf{O})$ estimates are given in Appendix E.

4.3 Word-Level Relationship

In this section, we test the relationship between reading time and informational quality at the *word level*. To do so, we fit linear mixed-effects models with reading time of an orthographic representation as the dependent variable and its IG as a fixed effect. We also included frequency, surprisal, contextual entropy, and (in English) word length as additional fixed effects, as well as by-subject and by-item random intercepts.

We find a significant effect of IG on reading time across all models and measures, with a consistent negative effect: the higher the informational quality of the input, the faster it is read. In Chinese, all three IG estimates were significant predictors of reading time: Qwen2.5-Zeroshot ($\beta = -7.53$ ms), Qwen2.5-Finetuned ($\beta = -10.19$ ms), and TransOCR ($\beta = -4.97$ ms). In English, the effects were even larger: Qwen2.5-Zeroshot ($\beta = -23.67$ ms), Qwen2.5-Finetuned ($\beta = -51.48$ ms), and TransOCR ($\beta = -66.42$ ms). All effects were statistically significant at $p < 0.001$.

4.4 Nonlinear Relationship Between Information Quality and Reading Time

While our linear regression models show that informational quality affects reading time, it makes (arguably strong) assumptions about the functional form of this relationship. In order to

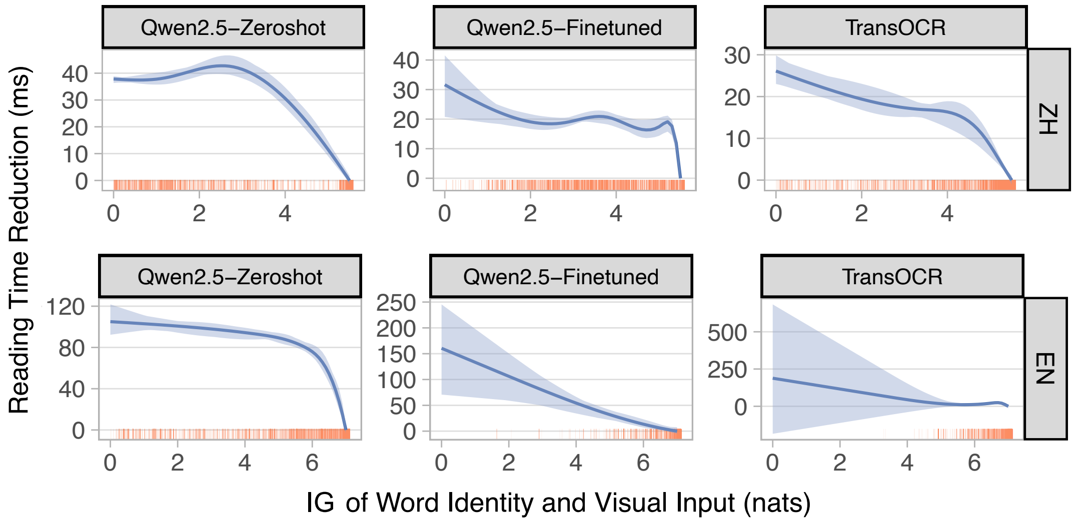

Figure 4: Relationship between informational quality of individual words (information gain; IG) and reading time slowdown. Solid blue lines are smoothed GAM fits; shaded regions show bootstrapped 95% confidence intervals. Red tick marks along the bottom (rug plots) indicate the distribution of IG data points. Reading times are aligned to end at zero at the highest MI end to emphasize relative reading time reductions.

get a better sense of how these two variables are related, we visualize them together in Figure 4. We used generalized additive models (GAMs). GAMs are models that allow for non-linear relationships between predictor and response variables. We fit GAMs to predict reading times with smooth terms for IG, controlling for frequency, surprisal, contextual entropy, and (for English) word length. We applied bootstrap smoothing over 20 resamples and computed confidence intervals for the estimated effects. We observe a consistent trend across both languages and all three models: reading time remains relatively stable at lower IG estimates but decreases rapidly as IG increases.

5 Discussion

Turning back to our main prediction, we argue that our results provide strong evidence that visual quality, as measured by mutual information, or information gain, impacts ease of processing. First, we find a consistent ordering, both in terms of reading times and mutual information, across our three experimental conditions. Second, we find a significant effect of the specific mutual information, or information gain, of individual words on reading times. While intuitive, the idea that bottom-up informational quality impacts ease of reading has not been quantified within a formal framework of reading. Our methods and experiments provide a specific estimate for the relationship between visual informational quality and reading times, which in English is between $25-66\, ms/bit$ and in Chinese $5-10\, ms/bit$. However, these numbers should be taken only as rough estimates, as the exact functional form may not be linear.

Turning now to our two sub research questions outlined in section 3.1: Interestingly, we find that information is not distributed evenly between the top and bottom half of words. Both English and Chinese place more information about word identity in the top half of their orthographic systems, a feature which we argue is reflected in the quicker reading times for our *Upper* condition. Interestingly, Pimentel et al. (2021) find similar informational asymmetries between the beginnings and ends of words, using an even wider set of languages. Exploring whether their asymmetry in reading times and extending our results to more languages is an important direction for future research. Finally, we find some suggestive evidence that this asymmetry is stronger in English, reflected in the larger effect sizes for the *Upper* vs. *Lower* contrast in our reading data. Future work should investigate such differences in greater detail.

6 Limitations

There are several limitations with the present work. In our formal model, we made two assumptions—that visual samples of a given word E are drawn i.i.d. during reading, and that visual inputs are conditionally independent from each other given W. These assumptions are strong, however, they are compatible with a "simple but fast" approach to reading. We discuss them in more detail in Appendix A.

Another limitation concerns our approach to estimating mutual information between word identity and orthographic representation in Chinese. We used characters, rather than lexical words, as the unit of analysis. This choice was motivated by two considerations: first, the average word length in our OneStopQA Chinese dataset is approximately 1.4 characters; second, Chinese characters, unlike English letters, carry substantial visual and semantic complexity. As such, characters may serve as a more suitable unit for modeling bottom-up visual processing in Chinese, analogous to words in English. Nonetheless, using lexical words might produce slightly different estimates of mutual information. Future work could examine whether similar patterns hold when words are used instead of characters.

One other limitation of the present work has to do with the languages studied. While we chose two languages that were topologically distinct, and used different types of orthographic systems, they represent only two language samples. Extending to more languages will be important to generalize the conclusions of this work.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Yevgeni Berzak, Jonathan Malmaud, and Roger Levy. 2020. STARC: Structured annotations for reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5726–5735, Online. Association for Computational Linguistics.

Klinton Bicknell and Roger Levy. 2010. A rational model of eye movement control in reading. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1168–1178.

Marc Brysbaert and Boris New. 2024. The subtlex word frequency norms. In *Reference Module in Social Sciences*. Elsevier.

Thomas M Cover. 1999. *Elements of information theory*. John Wiley & Sons.

Jon Gauthier and Roger Levy. 2023. The neural dynamics of word recognition and integration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 980–995, Singapore. Association for Computational Linguistics.

Edward Gibson, Leon Bergen, and Steven T. Piantadosi. 2013. Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, 110(20):8051–8056.

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*.

Gordon E Legge, Timothy S Klitz, and Bosco S Tjan. 1997. Mr. Chips: An ideal-observer model of reading. *Psychological Review*, 104(3):524.

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

Roger Levy. 2008b. A noisy-channel model of human sentence comprehension under uncertain input. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 234–243, Honolulu, Hawaii. Association for Computational Linguistics.

Zhecheng Li, Guoxian Song, Yujun Cai, Zhen Xiong, Junsong Yuan, and Yiwei Wang. 2025. Texture or semantics? vision-language models get lost in font recognition. *arXiv preprint arXiv:2503.23768*.

Dennis Norris. 2006. The Bayesian reader: Explaining word recognition as an optimal bayesian decision process. *Psychological Review*, 113(2):327.

Metehan Oğuz, Cui Ding, Ethan Gotlieb Wilcox, and Zuzanna Fuchs. 2025. Using MoTR to probe gender agreement in Russian.

Tiago Pimentel, Ryan Cotterell, and Brian Roark. 2021. Disambiguatory signals are stronger in word-initial positions. In *Proceedings of the 16th*

Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 31–41, Online. Association for Computational Linguistics.

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.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online. Association for Computational Linguistics.

Tamar I Regev, Chiebuka Ohams, Shaylee Xie, Lukas Wolf, Evelina Fedorenko, Alex Warstadt, Ethan G Wilcox, and Tiago Pimentel. 2025. The time scale of redundancy between prosody and linguistic context. *arXiv preprint arXiv:2503.11630*.

Claude E Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Robyn Speer. 2022. rspeer/wordfreq: v3.0.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Ethan Gotlieb Wilcox, Cui Ding, Giovanni Acampa, Tiago Pimentel, Alex Warstadt, and Tamar I Regev. 2025. Using information theory to characterize prosodic typology: The case of tone, pitch-accent and stress-accent. *arXiv preprint arXiv:2505.07659*.

Ethan Gotlieb Wilcox, Cui Ding, Mrinmaya Sachan, and Lena Ann Jäger. 2024. Mouse Tracking for Reading (MoTR): A new naturalistic incremental processing measurement tool. *Journal of Memory and Language*, 138:104534.

Lukas Wolf, Tiago Pimentel, Evelina Fedorenko, Ryan Cotterell, Alex Warstadt, Ethan Wilcox, and Tamar Regev. 2023. Quantifying the redundancy between prosody and text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9765–9784, Singapore. Association for Computational Linguistics.

Haiyang Yu, Xiaocong Wang, Ke Niu, Bin Li, and Xiangyang Xue. 2023. Scene text segmentation with text-focused transformers. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 2898–2907, New York, NY, USA. Association for Computing Machinery.

A Assumptions of Formal Model

In this appendix, we discuss our two assumptions about our samples of evidence, $\mathbf{E}$, namely that they are drawn i.i.d., and that they are conditionally independent of each other, given W. First, given these two assumptions, we walk through the step from 4a to 4b. First, we have by the definition of mutual information:

$$I(W; E_i \mid \mathbf{E}_{1:i-1}) \tag{11}$$

$$= \sum_{i=1}^{k} I(W; E_i \mid \mathbf{E}_{1:i-1}) \tag{12}$$

$$= \sum_{i=1}^{k} H(E_i \mid \mathbf{E}_{1:i-1}) - H(E_i \mid W, \mathbf{E}_{1:i-1}) \tag{13}$$

Assuming that the samples E are drawn independently of each other, we have, for the first term in this sum that $H(E \mid \mathbf{E}_{1:i-1}) = H(E)$. That is, the previous samples don't influence the entropy of the current sample. Furthermore, assuming conditional independence between the samples, given W, we have that $H(E \mid W, \mathbf{E}_{1:i-1}) = H(E \mid W)$. Therefore, we can rewrite as:

$$= \sum_{i=1}^{k} H(E_i) - H(E_i \mid W) \tag{14}$$

$$= \sum_{i=1}^{k} I(E_i; W) \tag{15}$$

which, given the symmetry of mutual information, is what we have in 4b.

Regarding our assumptions, the first one means that we model the reader as not making their decision about what to sample next based on information about previous samples within a given word. The second assumption means that if the reader knows the word's identity, then previous samples will not necessarily help them to predict what will be sampled next. We believe that both of these (especially the first one) are somewhat strong assumptions. However, they are compatible with the view that readers adopt a simple, but fast, sampling strategy, in which prior evidence from samples does not determine future sampling behavior. Given that reading happens at a very quick timescale, where word identification takes potentially only tens of milliseconds, such a "simple but fast" approach is not unreasonable.

B Qwen2.5-VL-7B-Instruct Fine-Tuning Details

We fine-tune Qwen2.5-VL-7B-Instruct using QLoRA with 4-bit quantization and LoRA adapters applied to attention projection layers with rank 8, $\alpha = 16$, and dropout 0.05. The model is trained for up to 100 epochs. Early stopping is applied based on validation loss. The training will terminate if no improvement for three consecutive epochs. AdamW (learning rate 2e-4), batch size 4, gradient accumulation of 8, and gradient clipping of 1.0. Training data consists of system and user prompts with bottom-half character images; the model predicts a single Chinese character. We formatted the input using Qwen's chat template and computed the loss on the assistant tokens. Image inputs are processed using the Qwen processor. Training is conducted on a single GPU (RTX 3090 Ti). Each training sample consists of a fixed system prompt and a task-specific user prompt. For example, for the lower-half recognition task, the templates used are as follows:

Chinese prompt

> **<system prompt>** 你是一个善于识别汉字的智能助手。图片只展示了一个汉字的下半部分，请你根据下半部分准确识别该汉字，只回答一个汉字。
>
> **<user prompt>** 这张图片显示的是一个汉字的下半部分，上半部分被遮挡住了。请根据可见部分判断这是什么汉字，只回答一个汉字，不要包含其他内容。这个汉字是：

English prompt

> **<system prompt>** You are a helpful assistant that can identify English words in images. The image will show only the lower half of an English word, with the upper half masked. Identify the word accurately based on the visible portion. Please answer with a single word, and do not include any other text.
>
> **<user prompt>** The image contains the lower half of an English word. The upper half is masked. What is the word in the image? Please answer with a single word, and do not include any other text. The word is:

C TransOCR Training Details

We trained the Transformer-based OCR model (TransOCR) for character recognition using the PyTorch framework. The model takes grayscale images resized to 32×256 pixels as input and is trained to predict character sequences in an autoregressive manner. Training was conducted using the Adadelta optimizer (ρ = 0.9, weight decay = 1e-4) with an initial learning rate of 1.0 and a batch size of 16. The loss function was standard cross-entropy over predicted character classes. We applied early stopping with a patience of 5 epochs based on validation accuracy.

All models were trained on two NVIDIA GPUs (RTX 3090 Ti) with multi-GPU support (DataParallel), and model checkpoints were saved at each epoch. The best-performing model was selected based on validation accuracy.

During inference, character predictions were generated step-by-step. At every step, the model outputs a probability distribution over the character vocabulary via a softmax layer. The conditional entropy is computed using the standard formula $H(\mathbf{p}) = -\sum_{i=1}^{N} p_i \log p_i$, where p_i is the predicted probability of the i-th character, given the input image.

D Self-Rated Ease of Reading

In both Chinese and English, participants overwhelmingly rated the upper half of words as easier to read. This asymmetry was more pronounced in English, where 91% of participants preferred the upper half, compared to 75% in Chinese.

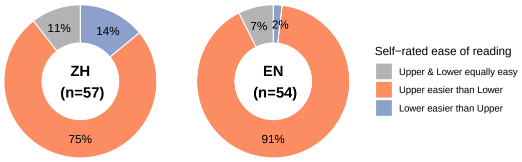

Figure 5: Self-rated ease of reading across visibility conditions. Participants were asked to judge whether the upper or lower half of words was easier to read.

E Mutual information estimates (nats)

Model	**Full**	**Upper**	**Lower**
Qwen2.5-Zeroshot	5.42	0.27	0.32
Qwen2.5-Finetuned	5.57	3.62	3.27
TransOCR	5.26	4.09	3.17

Table 1: Mutual information $I(W; \mathbf{O})$ in Chinese.

Model	**Full**	**Upper**	**Lower**
Qwen2.5-Zeroshot	6.99	5.74	3.86
Qwen2.5-Finetuned	7.11	7.01	6.66
TransOCR	7.07	7.00	6.68

Table 2: Mutual information $I(W; \mathbf{O})$ in English.