Where Patients Slow Down: Surprisal, Uncertainty, and Simplification in French Clinical Reading

Oksana Ivchenko[1]* **Alamgir Munir Qazi**[2]* **Jamal Abdul Nasir**[2]
[1]Univ. Lille, CNRS, UMR 8163 - STL, F-59000 Lille, France
[2]School of Computer Science, University of Galway, Ireland
{oksana.ivchenko.etu@univ-lille.fr}
{a.qazi1, jamal.nasir}@universityofgalway.ie

Abstract

This eye-tracking study links language-model surprisal and contextual entropy to how 23 non-expert adults read French health texts. Participants read seven texts (clinical case, medical, general), each available in an Original and Simplified version. Surprisal and entropy were computed with eight autoregressive models (82M–8B parameters), and four complementary eye-tracking measures were analyzed. Surprisal correlates positively with early reading measures, peaking in the smallest GPT-2 models ($r \approx 0.26$) and weakening with model size. Entropy shows the opposite pattern, with negative correlations strongest in the 7B-8B models ($r \approx -0.13$), consistent with a skim-when-uncertain strategy. Surprisal effects are largest in *Clinical Original* passages and drop by $\sim$20% after simplification, whereas entropy effects are stable across domain and version. These findings expose a scaling paradox – where different model sizes are optimal for different cognitive signals – and suggest that French plain-language editing should focus on rewriting high-surprisal passages to reduce processing difficulty, and on avoiding high-entropy contexts for critical information.

1 Introduction

Developing efficient methods to detect reading difficulty in healthcare materials is crucial for text simplification efforts (Fox, 2014). However, standard readability metrics provide limited insight into where and why readers struggle. Healthcare materials are frequently difficult for patients to understand (Rey et al., 2023), yet traditional measures fail to capture the localized nature of reading difficulty. Eye-tracking shows that effort is highly localized: readers invest extra time where their expectations are violated or where contextual uncertainty is high, then skim easier stretches (Ehrlich and Rayner, 1981; Rayner, 1998). Probabilistic language models (LMs) quantify these two information states that drive reading difficulty. Surprisal captures the unexpectedness of the word that actually appears and robustly predicts reading time (Smith and Levy, 2013; Goodkind and Bicknell, 2018). Contextual entropy captures an anticipatory state: high entropy can induce skipping or shorter gazes, whereas low entropy makes prediction errors costlier (Linzen and Jaeger, 2016; Pimentel et al., 2023). Early eye movement measures reflect immediate processing difficulty, while late measures indicate integration and comprehension costs (Camblin et al., 2007). Recent work reveals a scaling paradox: surprisals from very large transformers ($>$ 2B) can diverge from human reading times, whereas mid-sized GPT-2 models sometimes align better (Oh and Schuler, 2023). This suggests that model size alone does not guarantee better psycholinguistic validity. Moreover, nearly all evidence comes from English newspapers or novels, with minimal work on health genres or French.

Using French clinical and general texts in original and simplified versions, we investigate which LM-based predictors best track reading difficulty for automated simplification systems. Specifically:
RQ1: Do effects vary by Domain (Clinical vs General) and Version (Original vs Simplified)?
RQ2: Which LMs align best with human data for each predictor?

In what follows, we describe the corpus texts (original documents and the creation of their simplified versions) in Section 2. In Section 3, we present the methodology. Section 4 is dedicated to the results. Finally, we conclude in Section 5.

*These authors contributed equally to this work.

Proceedings of First International Workshop on Gaze Data and NLP associated with RANLP 2025, pages 52–57, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-104-2-007

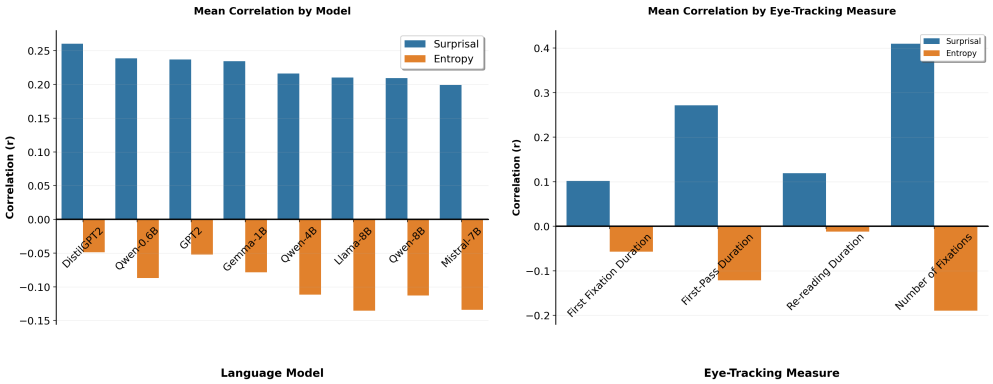

Figure 1: Mean Correlation

2 Data

2.1 Texts

We constructed a dataset of French-language texts from general and medical domains, based on excerpts from two corpora: CLEAR (Grabar and Cardon, 2018) and CAS (Grabar et al., 2020).

Each text was manually simplified following (OCDE, 2015) guidelines through syntactic, lexical, and semantic modifications, typically resulting in longer, clearer versions. Counterbalancing eliminated familiarity bias by exposing each participant to only one version of each text. Table 2 in the Appendix contains the full breakdown by words, sentences and screens.

2.2 Participants & Procedure

Gaze data were recorded using a Tobii Pro Spectrum eye tracker sampling at 600 Hz.

Texts were presented slide-by-slide, with some slides including comprehension questions for engagement. Tobii Pro Lab managed text presentation and automatically defined word-level Areas of Interest (AOIs).

The sample comprised 23 French participants aged 18–42 years (M = 22.8, SD = 6.2). Participants come from various social backgrounds - including students, doctoral students, and working professionals - but none have medical training.

3 Modeling

3.1 Language Models

We evaluated eight pre-trained autoregressive LMs spanning nearly three orders of magnitude in size (Table 1). Selection criteria were (i) good French coverage and (ii) architectural variety: four Byte-Pair Encoding (BPE) tokenisers (DistilGPT-2, GPT-2, two Qwen variants, Llama-3.1) and four SentencePiece models (Gemma-1B, Mistral-7B, Qwen-4B, Llama-8B). All models were run via `HuggingFace Transformers` with identical inference settings (`temperature` = 0, no sampling).

3.2 Eye-Movement Measures

We focus on four eye–movement measures, each indexing a distinct stage of processing:

Duration of first fixation (DFF) – immediate lexical access (time of the very first fixation);
First-pass duration (FPD) – initial comprehension (total dwell time during the first encounter);
Number of fixations (NFix) – overall processing effort (count of all fixations on the word);
Re-reading duration (RRD) – later integration/repair (time spent re-visiting the word).

These measures collectively span the complete timeline from initial word recognition to final comprehension, allowing us to assess how psycholinguistic predictions manifest across different aspects of the reading process.

3.2.1 Surprisal

We computed word-level surprisal as the negative log probability of each word given its left context:

$$\text{Surprisal}(w_i) = -\log_2 P(w_i \mid w_1, \ldots, w_{i-1}) \quad (1)$$

For each sentence, we obtained the model's probability distribution over the vocabulary at each position using a forward pass, extracted the probability assigned to the observed word, and converted to bits using base-2 logarithms. Surprisal values were aggregated from subword tokens to word level by

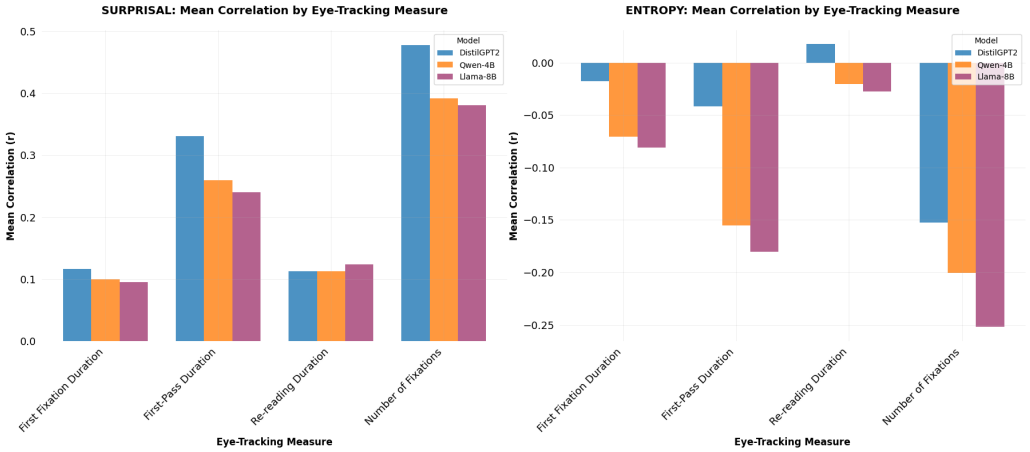

Figure 2: Mean correlations between model-generated surprisal (left panel) and entropy (right panel) with four eye-tracking measures. DistilGPT2 (82M parameters) consistently outperforms medium-sized Qwen-4B (4B parameters) and large Llama-8B (8B parameters) across the reading measures.

Model	Parameters
DistilGPT2	82M
GPT2	124M
Qwen3-0.6B	600M
Gemma-3-1B-IT	1B
Qwen3-4B	4B
Mistral-7B-Instruct-v0.3	7B
Qwen3-8B	8B
Llama-3.1-8B-Instruct	8B

Table 1: Overview of language models evaluated in this study, ranging from 82M to 8B parameters.

summing surprisal across all tokens comprising each word.

3.2.2 Contextual Entropy

We calculated the entropy of the model's predictive distribution at each word position:

$$H(i) = -\sum_{w} P(w \mid c_i) \log_2 P(w \mid c_i) \quad (2)$$

where $c_i = w_1, \ldots, w_{i-1}$ represents the left context.

This measure captures the model's uncertainty about what word should come next, independent of the actual word that appears. Higher entropy values indicate greater uncertainty in the model's predictions.

3.3 Data Processing and Token Alignment

3.3.1 Pre-processing

We rebuilt sentence strings by concatenating word tokens and normalising surrounding punctuation. For eye-movement data, duration metrics kept only positive values, whereas count metrics kept zeros but dropped negatives. Outliers were trimmed with measure-specific cut-offs: the upper 99 % for durations and the upper 95 % for counts. Analyses were run only when a cell contained at least ten valid observations, ensuring stable statistics.

3.3.2 Character-Position Mapping Algorithm

The technical challenge involved aligning model subword tokens with human word boundaries. French words often tokenize into multiple subwords (e.g., "L'obstétrique" → ["L'", "obsté", "trique"]), but humans process complete orthographic words.

Our alignment algorithm proceeded as follows:
(1) Extract character spans for each token using the tokenizer's offset mapping
(2) Define word boundaries from whitespace-delimited text
(3) For each word, identify all overlapping tokens using character position intersection
(4) Sum surprisal values of overlapping tokens to obtain word-level surprisal
(5) Average entropy values across tokens within each word
(6) Handle edge cases (partial overlaps, missing tokens) with fallback procedures

This method generalizes across tokenization schemes and languages, enabling consistent surprisal calculation regardless of subword segmentation. The algorithm successfully aligned tokens with word boundaries across all experimental conditions.

3.4 Statistical Analysis

3.4.1 Pearson Correlation Coefficient

We employed Pearson product-moment correlation as our primary statistical measure to quantify the linear relationship between language model predictions and human eye-movement behavior. The Pearson correlation coefficient r is defined as:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (3)$$

where x_i represents individual language model predictions (surprisal or entropy values), y_i represents corresponding eye-movement measures, $\bar{x}$ and $\bar{y}$ are sample means, and n is the number of word-level observations.

3.4.2 Correlation Analysis Framework

For every participant – text – metric cell we computed Pearson correlations between each predictor and the corresponding eye measure. The fully crossed design produced 23 participants $\times$ 8 texts $\times$ 4 metrics $\times$ 2 predictors = 1 472 correlation tests (counterbalancing included).

Surprisal correlation : r between word-level surprisal and the eye metric.

Entropy correlation : r between contextual entropy and the eye metric.

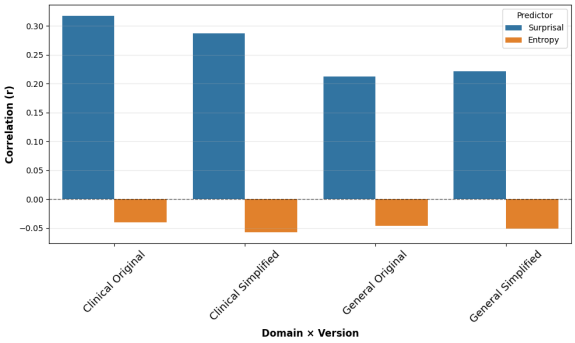

Figure 3: Domain Analysis

4 Results

Figure 1 summarises the aggregate correlations. In the left panel, surprisal (blue) is positive for every model, peaking in the two GPT-2 variants ($r \approx .25$) and tapering off as size increases. Entropy (orange) is negative and grows in magnitude, reaching $r \approx -0.14$ for the 7–8 B transformers. Hence small models best capture surprisal-driven slow-downs, while large models best capture the skim-when-uncertain effect indexed by entropy.

The right panel aggregates across models to compare eye-tracking metrics. Surprisal is strongest for *NFix* and *FPD*, weaker for *RRD*, and minimal for *DFF*. Entropy exhibits the reverse profile: it is most negative for *NFix*, moderate for *FPD*, and near zero on later measures – supporting the interpretation that surprisal indexes integration difficulty, whereas entropy reflects a strategic (skim-when-uncertain) allocation of attention.

Figure 2 contrasts a small (DistilGPT-2, 82 M), mid-size (Qwen-4B, 4 B) and large (Llama-8B, 8 B) model across the four eye metrics.

Surprisal. The ordering of effects is preserved across models, but magnitudes shrink as model size increases: DistilGPT-2 reaches $r = 0.48$ on *NFix* and $r = 0.33$ on *FPD*, whereas Llama-8B falls to $r = 0.38$ and $r = 0.24$, respectively. Small models therefore yield the clearest surprisal signal.

Entropy. The pattern is reversed. DistilGPT-2 shows near-zero correlations, Qwen-4B shows moderately negative correlations, and Llama-8B shows the strongest negative effects ($r = -0.25$ on *NFix*, $r = -0.18$ on *FPD*). The ranking of measures also flips: entropy effects are largest for fixation count and first-pass metrics, but minimal for *DFF* and *RRD*.

Figure 3 plots predictor strength by domain and simplification. Surprisal peaks in *Clinical Original* passages ($r \approx .32$), drops to $r \approx .29$ after simplification, and is lower overall in *General* texts ($r \approx .27 - .28$). Clinical terminology therefore amplifies error-driven slow-downs, and plain-language rewriting mitigates – but does not eliminate – this cost.

Entropy (orange) stays small and negative in every condition ($r \approx -0.03 - -0.05$) and shows no clear domain or version effect, implying that the skim-when-uncertain strategy is domain-invariant.

In short, simplification primarily reduces surprisal-based integration effort in specialized texts, while entropy-based allocation of attention remains unchanged.

5 Conclusion & Future Work

We demonstrate that LM-derived surprisal and entropy capture *different* aspects of French reading behavior, with effects that depend on text type: clinical originals produce the largest surprisal-driven slow-downs, while entropy effects remain modest

and stable across conditions. Small GPT-2 models best predict surprisal-based processing costs, whereas large 7–8B models best predict entropy-driven skimming behavior. Future work will (i) extend the corpus to longer passages and more readers, (ii) model text-level variation more explicitly by identifying which text properties modulate surprisal and entropy effects, (iii) investigate individual differences in reading strategies, and (iv) develop an automated simplification pipeline.

The current analysis is limited to clinical and general texts. Future studies will incorporate medical texts to examine domain effects more comprehensively.

Acknowledgments

This research is partially funded by MultiplEYE COST Action CA21131. Alamgir Munir Qazi is supported by the European Union's Horizon Europe programme under grant agreement No 101135757, project AI4Debunk[1].

This work was partially funded by the French National Agency for Research (ANR) as part of the CLEAR project (Communication, Literacy, Education, Accessibility, Readability), ANR-17-CE19-0016-01. Oksana Ivchenko is supported by the French National Agency for Research (ANR).

References

C. Christine Camblin, Peter C. Gordon, and Tamara Y. Swaab. 2007. The interplay of discourse congruence and lexical association during sentence processing: Evidence from ERPs and eye tracking. *Journal of Memory and Language*, 56(1):103–128.

Susan F. Ehrlich and Keith Rayner. 1981. Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, 20(6):641–655.

Susannah Fox. 2014. The social life of health information. Technical report, Pew Internet & American Life Project, Washington DC.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics*, pages 10–18, New Orleans, LA. Association for Computational Linguistics.

Natalia Grabar and Rémi Cardon. 2018. Clear – simple corpus for medical French. In *Workshop on Automatic Text Adaption (ATA)*, pages 1–11.

[1] https://ai4debunk.eu

Natalia Grabar, Clément Dalloux, and Vincent Claveau. 2020. CAS: corpus of clinical cases in French. *Journal of BioMedical Semantics*, 11(1):1–7.

Tal Linzen and T. Florian Jaeger. 2016. Uncertainty and expectation in sentence processing: Evidence from subcategorization probabilities. *Cognitive Science*, 40(6):1382–1411.

OCDE. 2015. *Guide de style de l'OCDE Troisième édition: Troisième édition*. OECD Publishing.

Byung-Doh Oh and William Schuler. 2023. Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1770–1784, Singapore. Association for Computational Linguistics.

Tiago Pimentel, Clara Meister, Ethan Wilcox, Roger P. Levy, and Ryan Cotterell. 2023. On the effect of anticipation on reading times. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 1234–1248, Toronto, Canada. Association for Computational Linguistics.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422.

Sylvie Rey, Aude Leduc, Xavier Debussche, Laurent Rigal, and Virginie Ringa. 2023. Une personne sur dix éprouve des difficultés de compréhension de l'information médicale. Études et Résultats 1269, Direction de la Recherche, des Études, de l'Évaluation et des Statistiques (DREES), Paris, France.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

6 Appendix

Table 2: Comparison of Original and Simplified Texts

text_type	text_name	version	total_screens	total_sentences	total_words
clinical	toxico	original	4	19	398
		simplified	5	29	469
clinical	gastro	original	3	13	255
		simplified	3	13	336
general	weekend	original	9	31	844
		simplified	9	49	811
general	camelot	original	8	42	840
		simplified	8	58	880
medical	obstetrics	original	12	57	1104
		simplified	12	65	1202
medical	stroke	original	3	10	276
		simplified	3	22	328
medical	ulcer	original	15	77	1526
		simplified	15	92	1551