Genre Matters: How Text Types Interact with Decoding Strategies and Lexical Predictors in Shaping Reading Behavior

Lena S. Bolliger, Lena A. Jäger
Department of Computational Linguistics, University of Zurich, Switzerland
{bolliger,jaeger}@cl.uzh.ch

Abstract

The type of a text profoundly shapes reading behavior, yet little is known about how different text types interact with word-level features and the properties of machine-generated texts and how these interactions influence how readers process language. In this study, we investigate how different text types affect eye movements during reading, how neural decoding strategies used to generate texts interact with text type, and how text types modulate the influence of word-level psycholinguistic features such as surprisal, word length, and lexical frequency. Leveraging EMTeC (Bolliger et al., 2025), the first eye-tracking corpus of LLM-generated texts across six text types and multiple decoding algorithms, we show that text type strongly modulates cognitive effort during reading, that psycholinguistic effects induced by word-level features vary systematically across genres, and that decoding strategies interact with text types to shape reading behavior. These findings offer insights into genre-specific cognitive processing and have implications for the human-centric design of AI-generated texts. Our code is publicly available at https://github.com/DiLi-Lab/Genre-Matters.

1 Introduction

The type or genre of a text influences the cognitive effort we expend at different stages of language processing (Blohm et al., 2022). A proxy for this cognitive load in language processing consists in the way we move our eyes during reading: not only do eye movements contain information about the properties and structure of the text being read, but they also provide insights into the cognitive mechanisms underlying language processing, as different words require a different amount of cognitive effort to be processed (Rayner, 1998; Rayner and Clifton, 2009). Given these qualities, eye movements have been leveraged to investigate readers' interactions with different text types, observing, for instance, that poetry leads to more regressions (Corcoran et al., 2023) or that fiction is read faster than nonfiction (Brysbaert, 2019). However, most of these studies have examined different genres in isolation and not directly pitted them against each other under the same experimental conditions, which would be crucial to make direct comparisons.

Moreover, while these studies do look at reading behavior in different text types, they do so in a coarse-grained manner by, for instance, considering overall reading time at the text level, thereby not accounting for word-level features which prompt reading patterns. These word-level features constitute well-established psycholinguistic phenomena. They include the *word length effect* (longer words take more time to read than shorter ones; Rayner, 2009; Hyönä and Olson, 1995; Just and Carpenter, 1980; Kliegl et al., 2004), the *lexical frequency effect* (frequent words are processed faster than infrequent ones; Forster and Chambers, 1973; Inhoff and Rayner, 1986), and the *surprisal effect* (high-surprisal words take longer to process than low-surprisal ones; Hale, 2001; Levy, 2008; Gruteke Klein et al., 2024; Xu et al., 2023). That these effects exist in different text genres has been corroborated extensively (Pimentel et al., 2023; Frank and Aumeistere, 2024; Kuperman et al., 2024; Torres et al., 2021, *i.a.*), but mainly in isolation. Examining how word-level features play out across text types, however, can reveal interactions between these features and text type properties and contribute insights to cognitive science by showing that the influence of certain psycholinguistic effects might be genre-dependent, such as that a reader's sensitivity to predictability in processing is a function of text type.

Recently, a growing body of research has examined the relationship between textual outputs by language models (LMs) and humans and whether there is similarity in language production or language understanding processes between the two (Venka-

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 7471–7488
November 4-9, 2025

traman et al., 2023; Giulianelli et al., 2023). An integral aspect of these textual outputs by the LMs is the decoding strategy used to generate the text and its alignment with cognitive processing strategies. So far, only one study (Bolliger et al., 2024) investigated how humans read texts generated by large language models (LLMs), focusing on how different models and decoding algorithms affect cognitive processing during reading and suggesting that decoding strategies can influence the linguistic properties of a text and, in turn, affect its readability. However, this line of work has not yet considered how these effects may interact with the type of text being generated. Investigating this interaction can highlight whether certain decoding methods are better suited, in terms of processing ease, for particular genres and can help ensure that AI systems generate texts in a way that aligns with our genre-specific processing strategies. The interplay between decoding method and genre-specific properties is thus an important but underexplored area.

This study investigates the effect of text type on reading behavior and its interaction with psycholinguistic phenomena as well as with neural decoding algorithms by tackling the following questions:

- RQ_1 Do different text types elicit different reading patterns, reflecting different cognitive demands during reading?
- RQ_2 Do well-established word-level predictors of reading behavior, such as surprisal, word length, and lexical frequency, interact with text type in shaping how people read?
- RQ_3 Do the neural decoding strategies used to generate texts of different text types interact with those text types in shaping reading behavior?

We consider these research questions exploratory and therefore refrain from formulating concrete hypotheses. To investigate these research questions, we leverage the **E**ye Movements on **M**achine-Generated **Te**xts **C**orpus (EMTeC, Bolliger et al., 2025), the first dataset containing eye-tracking data on LLM-generated texts across six different text types, generated using a variety of decoding algorithms. This dataset does not only allow for a direct comparison of reading behavior across different text types and how psycholinguistic effects vary between them, but also how they interact with decoding algorithms.

Our findings suggest that text type exerts a strong influence on cognitive effort during reading, that the magnitude of the psycholinguistic effects elicited by lexical features is modulated by text type, and that the decoding strategies used by language models interact with text types to shape the ease of processing machine-generated texts.

2 Related Work

Text type or genre has long been recognized as a key factor in shaping reading behavior. Poetry, for example, induces longer fixations and more regressions due to its atypical syntax, ambiguity, and foregrounded language (Blohm et al., 2022; Corcoran et al., 2023), and readers' eye movements differ even when identical content is presented in poetic versus prosaic layout (Fechino et al., 2020). In contrast, narrative fiction elicits more linear reading patterns, attributed to its predictability (Graesser et al., 2003). Studies comparing fiction and nonfiction suggest that fiction is read more quickly, a difference largely driven by word length and lexical complexity (Brysbaert, 2019; Corcoran et al., 2023). While these studies demonstrate genre-specific reading patterns, they typically examine one genre at a time, under differing experimental conditions, thereby limiting comparability. Our work fills this gap by comparing six genres directly within a controlled, unified dataset.

In parallel, a large body of work has investigated psycholinguistic predictors of reading difficulty, such as surprisal (Hale, 2001; Levy, 2008; Gruteke Klein et al., 2024; Xu et al., 2023; Shain et al., 2024, *i.a.*), word length (Rayner, 1998, 2009; Hyönä and Olson, 1995; Just and Carpenter, 1980; Kliegl et al., 2004; Gerth and Festman, 2021; Kuperman et al., 2024, *i.a.*), and lexical frequency (Forster and Chambers, 1973; Inhoff and Rayner, 1986; Chen and Ko, 2011; Torres et al., 2021, *i.a.*). These effects have been consistently observed across a wide range of genres, including narrative (Luke and Christianson, 2016, 2018; Cop et al., 2017; Salicchi et al., 2023; Frank and Aumeistere, 2024), expository (Kennedy et al., 2003; Xu et al., 2023; Goodkind and Bicknell, 2018), and scientific texts (Klein et al., 2025; Jakobi et al., 2025). Even stylistic deviations such as foregrounding in literary texts modulate these effects (Van den Hoven et al., 2016). Although these findings highlight the robustness of psycholinguistic predictors, few studies have investigated whether their magnitude or nature differs across

text types. Our study addresses this by systematically analyzing interactions between genre and word-level predictors within the same experimental setup.

Finally, recent research has begun examining how texts generated by large language models are processed by human readers. Bolliger et al. (2024) showed that decoding strategies, such as top-p sampling or greedy decoding, can influence reading behavior, although no single strategy consistently outperformed others across measures or models in terms of processing ease. Other studies have explored the structure and information distribution of LLM outputs from the perspective of predictability or information density (Giulianelli et al., 2023; Venkatraman et al., 2023), but these analyses were conducted at the sentence level and did not incorporate eye-tracking data or account for text type. To date, no study has examined whether and how the impact of decoding strategies interacts with word-level features. Our study fills this gap by leveraging EMTeC (Bolliger et al., 2025), which combines multiple genres, multiple decoding strategies, and human eye-tracking data.

3 Experiments[1]

3.1 Data

EMTeC We employ reading data from the **E**ye Movements on **M**achine-Generated **Te**xts **C**orpus (EMTeC, Bolliger et al., 2025), an English eye-tracking-while reading corpus whose stimuli were created with three different large language models (LLMs) — Phi-2 (Javaheripi et al., 2023), Mistral 7B Instruct (Jiang et al., 2023), and WizardLM (Xu et al., 2024) — using five decoding algorithms — greedy search, beam search, ancestral sampling, top-k sampling (Fan et al., 2018), and top-p sampling (Holtzman et al., 2020). The generated stimuli belong to six different types of text categories: *Non-fiction*, where the models were prompted to either write a description or an argumentation; *Fiction*, where the LLMs were instructed to write a short story or a dialogue between two characters; *Poetry*, where the LLMs were prompted to write a poem; *Summarization*, where they were asked to summarize an input text; *Article*, where they ought to craft a news article out of an article synopsis; and *Key-word text*, where the LLMs had to create texts based on a range of input key words.

Reading Measures We consider the binary reading measures (RMs) *fixated* (Fix; whether or not a word was fixated) and *first-pass regression* (FPReg; whether or not a regression was initiated in the first-pass reading of the word) and the continuous RMs *total fixation time* (TFT; the sum of all fixations on a word), *first-pass reading time* (FPRT; the sum of the durations of all first-pass fixations on a word), *re-reading time* (RRT; the sum of the durations of all fixations on a word that do not belong to the first pass), and *regression path duration* (RPD; the sum of all fixation durations starting from the first first-pass fixation on a word until fixating a word to the right of this word). While *total fixation time* and *fixated* indicate global language processing, *first-pass reading time* and *first-pass regression* indicate early and *re-reading time* and *regression path duration* late stages of processing. The continuous reading measures are log-transformed. For the reasoning behind the log-transformation of reading measures, please refer to Appendix A.

3.2 Predictors

Word-level features. We include word-level predictors, namely surprisal, lexical frequency, and word length, whose impact on eye movement behavior in reading is well-established and key to psycholinguistic theories of reading and, more broadly, language comprehension (Reichle et al., 2003; Engbert et al., 2005; Veldre et al., 2020; Rabe et al., 2024). *Surprisal* quantifies the predictability of a word. It is based on surprisal theory (Hale, 2001; Levy, 2008), which operationalizes the relationship between cognitive processing effort and word predictability and posits that the cognitive effort needed to process a word is a linear function of that word's predictability. More specifically, surprisal is the negative log-probability of a word conditioned on its preceding (linguistic and extra-linguistic) context. This quantity is approximated by a neural language model p_ϕ, which only takes the preceding linguistic context into account. As such, given a vocabulary Σ, the surprisal s of a word $w \in \Sigma$ at position t is defined as

$$s(w_t) := -\log_2 p_\phi(w_t \mid \boldsymbol{w}_{<t}), \qquad (1)$$

where $p_\phi(\cdot \mid \boldsymbol{w}_{<t})$ is the language model's approximate distribution of the true distribution $p(\cdot \mid \boldsymbol{w}_{<t})$ over words $w \in \Sigma$ in context $\boldsymbol{w}_{<t}$.[2] In the follow-

[1] Our code is available at https://github.com/DiLi-Lab/Genre-Matters

[2] That means surprisal is computed across sentence boundaries.

ing, surprisal is estimated with GPT-2 *small* (Radford et al., 2019), which has been shown to have the highest predictive power on reading times among LMs (Shain et al., 2024). As the reading measures are computed on the level of white-space separated words but LMs use tokenizers that separate words into sub-word tokens (Sennrich et al., 2016; Song et al., 2021), we aggregate surprisal to the word level by summing up the surprisal values of the individual sub-word tokens.[3] The *lexical frequency* of a word is the Zipf frequency obtained from the `wordfreq` library,[4] which presents the frequency of a word on a logarithmic scale[5] and is the word's base-10 logarithm of the number of times it appears in a billion words. *Word length* refers to the number of characters of a white space-delimited word, including adjacent punctuation.

Contrast Coding of Text Type and Decoding Strategy Both the factor `text type`, consisting of the levels *non-fiction, fiction, poetry, summarization, article,* and *key-word text*, as well as the factor `decoding strategy`, consisting of the levels *beam search, ancestral sampling, top-k sampling, top-p sampling,* and *greedy search*, are sum-contrast coded. Sum-contrast coding compares the dependent variable — the reading measure — for each but one level of the factor to the grand mean across all levels. That is, for a factor with k levels, it generates $k-1$ contrast variables. The levels *key-word text* and *greedy search* serve as the reference levels for `text type` and `decoding strategy`, respectively, and are only implicitly represented in the grand mean intercept. The comparisons are factor level minus grand mean. The factor levels are coded as 1, the grand mean as -1, meaning that the respective model coefficient represents the estimated difference from the grand mean associated with this factor level. A positive coefficient indicates that the dependent variable is higher (*i.e.*, increased processing effort) compared to the grand mean baseline. A more detailed description of contrast coding and sum contrasts as well as the contrast matrices can be found in Appendix D.

3.3 Methods

For the analyses, we utilize linear mixed-effects models: linear regressions for continuous variables and logistic regressions for binary variables. The linear model is defined as $\eta = X\boldsymbol{\beta} + Z\mathbf{b}$, where $X \in \mathbb{R}^{N\times P}$ is the fixed-effects design matrix including the intercept term, surprisal s, the z-score standardized lexical Zipf frequency f, word length l, and the sum contrast-coded factors `text type` tt, `decoding strategy` dec, and `model` m (to control for the effect of the LLM with which the texts were generated in EMTeC). The sample size is denoted by N, the number of predictors by P, and the number of subjects by J. The random-effects design matrix $Z \in \mathbb{R}^{N\times J}$ specifies a by-subject random intercept,[6] and the random intercepts $\mathbf{b} \in \mathbb{R}^J$ are assumed to follow $\mathbf{b} \sim \mathcal{N}(\mathbf{0}, \Sigma_\theta)$, with covariance matrix Σ_θ parametrized by the variance components θ. Conditional on the fixed effects $\boldsymbol{\beta} \in \mathbb{R}^P$ and random effects $\mathbf{b} \in \mathbb{R}^J$, the responses follow a generalized linear model determined by the response type Y: for continuous reading measures, $(Y \mid \boldsymbol{\beta}, \mathbf{b}) \sim \mathcal{N}(\eta, \sigma^2 I)$, and for binary reading measures, $(Y \mid \boldsymbol{\beta}, \mathbf{b}) \sim \text{Bernoulli}(\text{logit}^{-1}(\eta))$. We fit all models using the R library `lme4` (Bates et al., 2015). Statistical significance is determined with Satterthwaite's approximation (Satterthwaite, 1946) from the `lmerTest` library (Kuznetsova et al., 2017) for linear regressions and with Wald z-tests (Wald, 1943) for logistic regressions. Further details are provided in Appendix B.

3.4 RQ_1: The Effect of Text Types

To examine whether text type influences reading behavior overall, *i.e.*, across all decoding strategies, we fit a regression model of the form described above. Here, the fixed-effects design matrix $X \in \mathbb{R}^{N\times P}$ includes an intercept term, word length l, lexical frequency f, surprisal s, the five sum contrast-coded contrasts from the `text type` tt, and the two sum contrast-coded effects from `model` m, with corresponding coefficient vector $\boldsymbol{\beta} \in \mathbb{R}^P$. The random-effects design matrix $Z \in \mathbb{R}^{N\times J}$ specifies a by-subject random intercept, with $\mathbf{b} \in \mathbb{R}^J$ denoting the subject-specific intercept. The model is fitted on the data across all decoding strategies.

Results Figure 1 depicts the effect estimates of the sum-contrast coded text types on the prediction of the different reading measures. The reading pattern elicited by the different text types is mostly consistent across the different reading mea-

[3] For elaborations on the pooling of sub-word token surprisal values, refer to Appendix C.

[4] https://pypi.org/project/wordfreq/

[5] There exists a linear relationship between log-frequency and reading times.

[6] We do not include random effects for items, as the number of unique items is too low.

sures and the effects are mostly significant, even when controlling for the psycholinguistic covariates surprisal, word length, and lexical frequency. Poetry exhibits the strongest positive effects: readers spend more time overall reading words in poems; they spend more time in first-pass reading as well as in re-reading, and poetry induces more first-pass regressions as well as number of fixations on words. Fiction and non-fiction, on the other hand, show the strongest negative effects: they cause significantly fewer fixations and first-pass regressions and lower reading times at any stage of processing (total fixation times, first-pass reading times, and re-reading times). Summarization texts and articles are both close to average, although summarization texts cause slightly more-than-average fixations and first-pass regressions, while articles cause slightly less.

3.5 RQ_2: The Interaction between Word-Level Features and Text Types

In order to investigate how the psycholinguistic predictors surprisal, word length, and lexical frequency interact with text type to influence reading behavior across different measures, and to assess whether the strength of these linguistic effects changes depending on text type, we fit a regression model of the form described above. However, the fixed-effects design matrix $X \in \mathbb{R}^{N \times P}$ additionally includes—next to the intercept, surprisal s, lexical frequency f, word length l, the two sum contrast-coded effects of `model` m, and the five sum contrast-coded effects of `text type` tt—now also the interaction between each of word length l, lexical frequency f, surprisal s, with the five sum contrast-coded effects of `text type` tt.

Results The main effects of word length, lexical frequency, and surprisal serve as a sanity check: they are as expected and are plotted in Appendix E.

Figure 2 depicts the interaction effects of sum-contrast coded text types with the psycholinguistic predictors—word length, lexical frequency, and surprisal—and reveals nuanced patterns. In summarization texts, surprisal effects are stronger than average for early binary measures (fixations and first-pass regressions) but weaker for early and late reading times (first-pass reading times and regression-path durations), while lexical frequency effects are generally smaller than average. In poetry, the interactions with surprisal indicates a smaller-than-average effect of surprisal on first-pass and total fixation times but an above-average effect on regression paths. Lexical frequency effects are amplified during first-pass and re-reading times in poetry, and word length exerted stronger effects on fixation probability, total fixation and re-reading times. For non-fiction, surprisal has a stronger effect on fixation probability and first-pass reading time, while lexical frequency and word length effects are weaker. Fiction texts amplify lexical frequency effects across almost all reading measures, with high-frequency words particularly facilitating faster reading, and exhibit reduced word length effects except for first-pass regressions. Finally, article texts show stronger surprisal effects on total fixation and re-reading times, stronger word length effects, and mixed frequency effects.

3.6 RQ_3: The Interaction Between Decoding Strategies and Text Types

In order to assess how the different decoding strategies used to generate the texts and the text types that the LLMs were prompted to generate interact in influencing human reading behavior, we fit a regression model of the form described in Section 3.3. The fixed-effects design matrix $X \in \mathbb{R}^{N \times P}$ includes the sum contrast-coded effects of `decoding strategy` dec, as well as all interactions between the five sum contrast-coded contrasts of `text type` tt and the four sum contrast-coded contrasts of `decoding strategy` dec, in addition to the intercept term, surprisal s, lexical frequency f, word length f, the five sum contrast-coded effects of `text type` tt, and the two sum contrast-coded effects of `model` m.

Results The fixed effects of the psycholinguistic predictors are plotted in Appendix F as a sanity check and are as expected for the psycholinguistic predictors and the text types. The main effects of the decoding strategies are mostly not significantly different from the grand mean.

Figure 3 shows the interaction effects between text type and decoding strategy. For poetry, texts generated with ancestral sampling and top-k sampling exhibited shorter first-pass reading times, shorter regression path durations, and lower total fixation times compared to the grand mean, while texts generated with top-p decoding exhibited longer regression paths and higher total fixation times. For fiction, beam search was associated with fewer fixations and reduced re-reading times, whereas top-p decoding increased fixation proba-

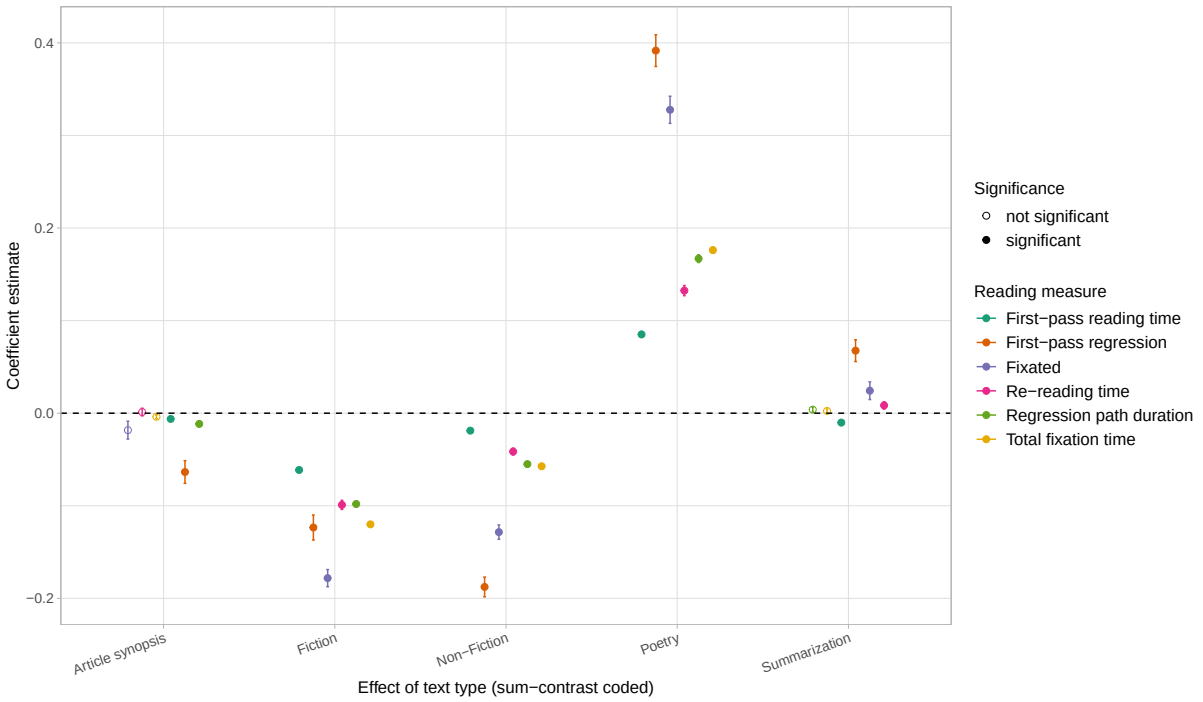

Figure 1: Effect estimates (mean and 95% CI) of sum-contrast coded text types on the prediction of different reading measures. Filled dots indicate that the effect is significantly different from the grand mean.

bility and sampling increased re-reading times. In non-fiction texts, top-p decoding was associated with fewer first-pass regressions, fewer fixations, and shorter total fixation times, while top-k decoding was associated with longer total fixation times. In summarization texts, top-k decoding was associated with fewer first-pass regressions, shorter regression paths, lower re-reading times, and reduced total fixation times, whereas beam search and sampling were associated with increased re-reading times and total fixation times. For articles, top-k decoding was associated with increased fixation probability, longer re-reading times, and higher total fixation times, while top-p decoding was associated with shorter total fixation times.

4 Discussion

The experimental results presented in this study contribute to the understanding of how *text types* influence reading behavior and how they *interact* with an *LLM's decoding strategy* and well-established *psycholinguistic phenomena* such as a word's predictability.

RQ_1: Genre-Driven Reading Patterns Directly pitting the different text types against each other under the same experimental conditions allows for making well-founded comparisons, and the findings for RQ_1 clearly demonstrate genre-driven divergences in reading behavior. Poetry emerged as the genre associated with the highest cognitive load across all stages of reading. This aligns with psycholinguistic theories that poetry's unconventional syntax and dense metaphoric context demand deeper interpretative processing and frequent re-analysis (Blohm et al., 2022; Corcoran et al., 2023; Fechino et al., 2020). Conversely, fiction and non-fiction texts were associated with significantly reduced cognitive demands, which suggests that narrative and expository prose align with readers' genre expectations and facilitate fluent reading (Graesser et al., 2003). Moreover, while summarization texts demand less cognitive effort to be processed than poetry, they demand significantly more cognitive effort than fiction and non-fiction. These findings demonstrate that the properties of different genres profoundly shape real-time cognitive processing during reading. They also underscore that poetry remains cognitively unique among genres — a pattern that persisted even though the stimuli were machine-generated, highlighting the robustness of genre-specific processing strategies.

RQ_2: Psycholinguistic Predictors Interact with Genre The genre-specificity in reading behavior is further corroborated and expanded upon in the results of the experiment answering RQ_2,

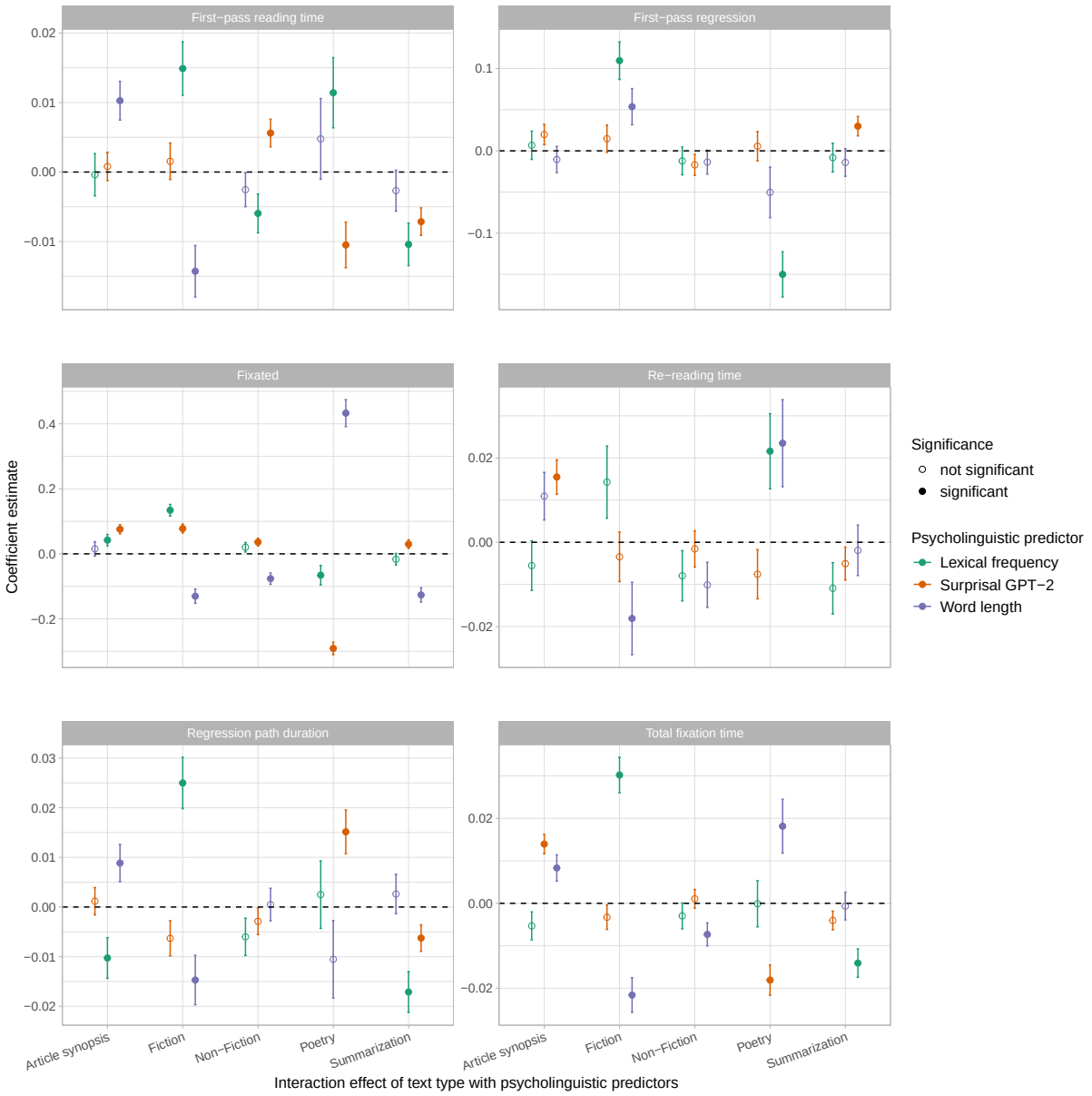

Figure 2: Interaction effects (mean and 95% CI) between text types and word-level predictors. A filled dot indicates that the interaction is significant ($p < 0.05$).

which investigates whether the word-level predictors of reading behavior surprisal, word length, and lexical frequency interact with text type. In poetry, surprisal had a weaker-than-average effect on early reading measures (FPRT), but a stronger-than-average effect on regression paths. This implies the following: readers tolerate local unpredictability in poetry during initial reading, possibly because they already anticipate the unpredictability, but they experience delayed integration difficulties that require re-reading and re-evaluation. Fictional texts demonstrated a higher-than-average effect of lexical frequency. They also exhibited a lower-than-average word length effect, which indicates that readers' sensitivity to word length is reduced compared to other text types. In non-fiction, surprisal effects on fixation probability and FPRTs were heightened, while lexical frequency and word length effects were weaker: readers seem to engage more heavily with predictive mechanisms during informational text reading, possibly due to the structured, factual nature of the content. In articles, surprisal effects were also stronger than on average, also indicating that the informative nature of the text makes readers engage in predictive processing. However, in contrast to non-fiction, this interaction between surprisal and genre was observed not in first-pass measures, but rather in re-reading time

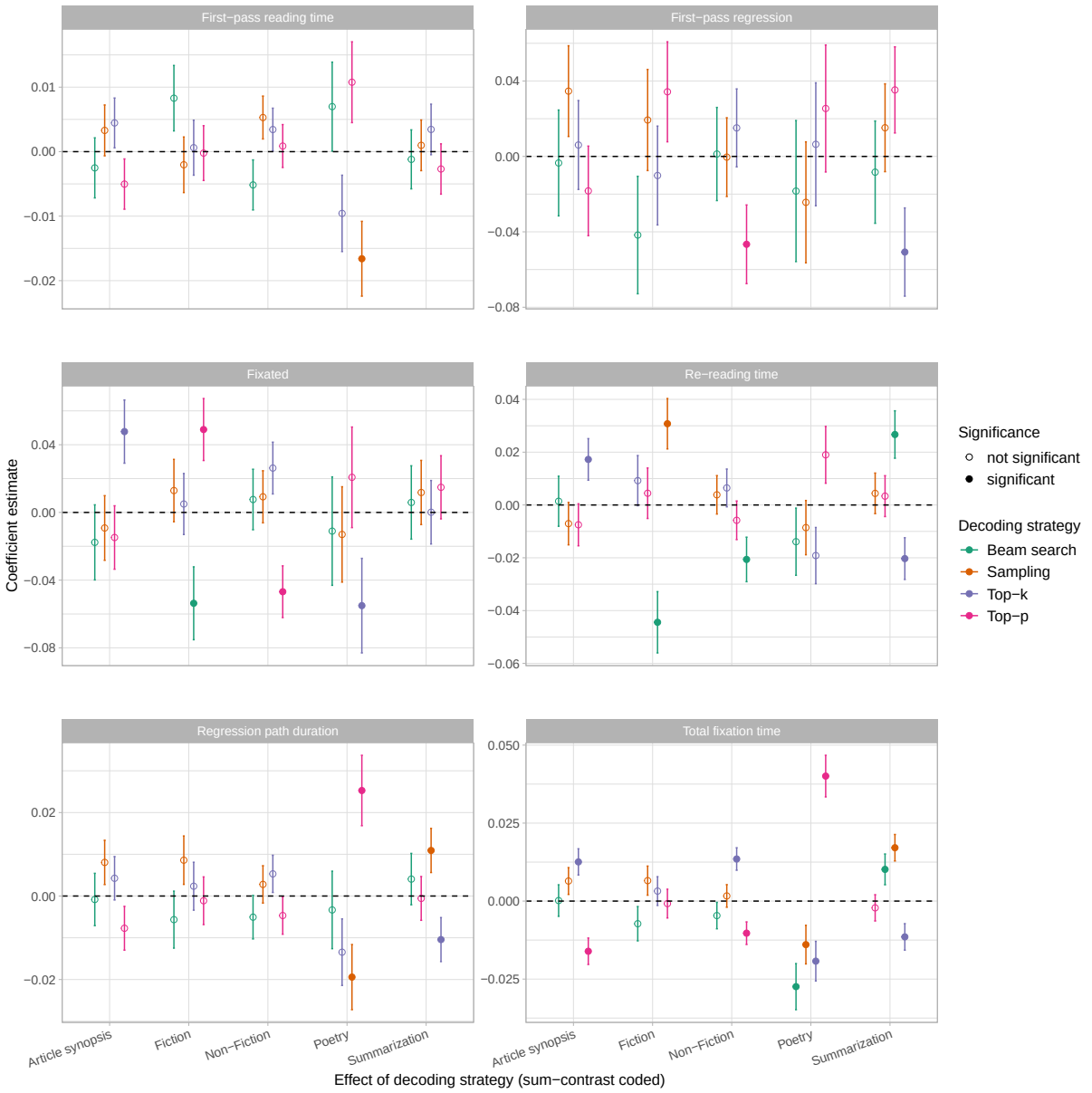

Figure 3: Interaction effects (mean and 95% CI) between sum-contrast coded decoding strategy and text type. A filled dot indicates that the interaction is statistically significant, meaning that the effect of a combination of text type and decoding strategy significantly differs from the effect predicted by the additive fixed effects alone.

and total fixation time. Articles also induce higher-than-average word length effects and lower-than-average lexical frequency effects, which might be interpreted as evidence for relying more heavily on decoding at the sub-word level (e.g., grapheme or syllable level) than on holistic lexical access.

These findings underline that while classic psycholinguistic predictors like surprisal, lexical frequency, and word length remain robust across genres, the magnitude and timing of their effects vary systematically with text type. Readers dynamically adapt their cognitive strategies depending on genre-specific expectations and structures.

RQ$_3$: Interaction between Decoding Strategy and Genre We further found that while the main effects of the decoding strategies utilized to generate the texts were minimal, their interactions with genre revealed meaningful patterns. In poetry, texts generated with the sampling-based strategies ancestral sampling and top-k sampling were easier to process — yielding shorter FPRTs, shorter RPDs, and lower TFTs — compared to those generated with top-p sampling, which paradoxically increased cognitive effort. This goes against the intuition that poetry generated with stochastic strategies requires more cognitive effort to be processed. This find-

ing suggests that moderate stochasticity benefits poetry by fostering the unpredictability and variability that readers expect, whereas the specific distribution of probabilities under top-p sampling may have introduced irregularities detrimental to coherent interpretation. In fiction, deterministic decoding via beam search facilitated the reading experience, reducing fixation probability and rereading times, whereas stochastic decoding strategies (sampling, top-p) introduced mild disruptions. This aligns with the intuition that narratives benefit from high predictability and coherence. In nonfiction, moderate randomness introduced by top-p decoding surprisingly facilitated reading — reducing regressions, fixations, and TFTs — while top-k decoding complicated it. This finding suggests that informational texts may benefit from slight variability, which might enhance engagement without compromising clarity. In summarization texts, top-k decoding led to the easiest reading (fewer regressions, shorter reading times), while both beam search and sampling complicated processing. This is intriguing because one might expect beam search to yield clear, coherent summaries — highlighting that stochastic decoding may sometimes better balance informativeness and readability. For articles, top-k decoding increased cognitive load, while top-p decoding decreased it, again emphasizing that subtle differences in decoding randomness can have substantial cognitive effects depending on genre.

In sum, these results demonstrate that no single decoding strategy universally optimizes readability. Rather, the ideal decoding method is crucially dependent on the genre and its associated cognitive demands as well as genre-specific expectations. This has direct implications for the design of human-centric LLM applications: depending on the desired use case or target population, generation systems may adapt decoding strategies to optimize user comprehension by facilitating reading ease, thereby matching the desired properties of different text types.

Broader Implications for Cognitive Science Overall, our findings have important implications for both cognitive science and AI research. From a cognitive perspective, the study reinforces the view that genre deeply shapes cognitive processing strategies during reading. Not only does it affect the baseline ease or difficulty of reading, but it also modulates the impact of core psycholinguistic variables like surprisal, lexical frequency, and word length. These results imply that cognitive models of reading must account for genre as a systematic source of variance, not merely as a surface-level property.

Implications for AI and NLP From an AI and NLP perspective, our results highlight that *how* a text is generated matters just as much as *what* genre it is intended to emulate. Different decoding strategies differentially align with text types in terms of ease of processing, affecting the cognitive accessibility of LLM-generated texts. As LLMs increasingly generate content for educational, journalistic, and entertainment purposes, understanding and optimizing for genre-appropriate readability will be crucial.

Methodological Contributions Finally, studying AI-generated texts provides a new lens through which to test cognitive theories: by controlling genre and text structure via generation parameters, we can probe the flexibility and robustness of human reading strategies in a way that complements traditional studies on human-written texts.

5 Conclusion

This study shows that text type significantly shapes reading behavior, modulating not only overall cognitive demands but also the strength and manifestation of core psycholinguistic effects. Genres like poetry induce higher effort, while fiction and nonfiction support easier processing. We further find that decoding strategies interact with genre in nontrivial ways, indicating that optimizing readability in machine-generated texts requires genre-sensitive approaches. These results highlight the need for adaptive generation systems that align with genre-specific cognitive norms.

Limitations

Several limitations must be acknowledged. First, while EMTeC provides a unique opportunity to study eye movements across machine-generated texts of different types, it does not include human-written baselines, which limits direct comparisons between human and machine text processing. Second, the texts were generated using only three LLMs and five decoding strategies, which may not capture the full diversity of possible outputs or decoding configurations. Third, the study focuses on adult readers and English texts; results may not generalize to different age groups, languages, or literacy backgrounds. Finally, while we account for core psycholinguistic predictors, other linguistic variables such as syntactic complexity or discourse coherence were not directly controlled and could influence reading behavior. Finally, the stimuli in EMTeC are not representative samples of their respective genres as a whole, as eye-tracking studies require constrained stimulus sets that do not support broad genre coverage.

Acknowledgments

We thank Cui Ding for her help with formulating the methodology and model formalizations. This work was supported by the COST Action MultiplEYE, CA21131.

References

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Stefan Blohm, Stefano Versace, Sanja Methner, Valentin Wagner, Matthias Schlesewsky, and Winfried Menninghaus and. 2022. Reading poetry and prose: Eye movements and acoustic evidence. *Discourse Processes*, 59(3):159–183.

Lena S. Bolliger, Patrick Haller, Isabelle C. R. Cretton, David R. Reich, Tannon Kew, and Lena A. Jäger. 2025. EMTeC: A corpus of eye movements on machine-generated texts. *Behavior Research Methods*, 57(7):189.

Lena S. Bolliger, Patrick Haller, and Lena A. Jäger. 2024. On the alignment of LM language generation and human language comprehension. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 217–231, Miami, Florida, US. Association for Computational Linguistics.

Marc Brysbaert. 2019. How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of Memory and Language*, 109:104047.

Minglei Chen and Hwawei Ko. 2011. Exploring the eye-movement patterns as Chinese children read texts: a developmental perspective. *Journal of Research in Reading*, 34(2):232–246.

Uschi Cop, Nicolas Dirix, Denis Drieghe, and Wouter Duyck. 2017. Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, 49:602–615.

Rhiannon Corcoran, Christophe de Bezenac, and Philip Davis. 2023. 'looking before and after': Can simple eye tracking patterns distinguish poetic from prosaic texts? *Frontiers in Psychology*, 14:1066303.

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

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Marion Fechino, Arthur M Jacobs, and Jana Lüdtke. 2020. Following in Jakobson and Lévi-Strauss' footsteps: A neurocognitive poetics investigation of eye movements during the reading of baudelaire's 'les chats'. *Journal of Eye Movement Research*, 13(3):10–16910.

Kenneth I. Forster and Susan M. Chambers. 1973. Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior*, 12(6):627–635.

Stefan L. Frank and Anna Aumeistere. 2024. An eye-tracking-with-EEG coregistration corpus of narrative sentences. *Language Resources and Evaluation*, 58(2):641–657.

Sabrina Gerth and Julia Festman. 2021. Reading development, word length and frequency effects: An eye-tracking study with slow and fast readers. *Frontiers in Communication*, 6:743113.

Mario Giulianelli, Sarenne Wallbridge, and Raquel Fernández. 2023. Information value: Measuring utterance predictability as distance from plausible alternatives. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5633–5653, Singapore. Association for Computational Linguistics.

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 (CMCL 2018)*, pages 10–18, Salt Lake City, Utah. Association for Computational Linguistics.

Arthur C. Graesser, Danielle S. McNamara, and Max M. Louwerse. 2003. What do readers need to learn in order to process coherence relations in narrative and expository text. *Rethinking Reading Comprehension*, 82:98.

Keren Gruteke Klein, Yoav Meiri, Omer Shubi, and Yevgeni Berzak. 2024. The effect of surprisal on reading times in information seeking and repeated reading. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 219–230, Miami, FL, USA. Association for Computational Linguistics.

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*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations (ICLR 2020)*.

Jukka Hyönä and Richard K Olson. 1995. Eye fixation patterns among dyslexic and normal readers: effects of word length and word frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(6):1430.

Albrecht Werner Inhoff and Keith Rayner. 1986. Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception & Psychophysics*, 40(6):431–439.

Deborah N. Jakobi, Thomas Kern, David R. Reich, Patrick Haller, and Lena A. Jäger. 2025. PoTeC: A German naturalistic eye-tracking-while-reading corpus. *Behavior Research Methods*, 57(8):211.

Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Sebastien Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, Suriya Gunasekar, Mojan Javaheripi, Piero Kauffmann, Yin Tat Lee, Yuanzhi Li, Anh Nguyen, Gustavo de Rosa, Olli Saarikivi, Adil Salim, and 9 others. 2023. Phi-2: The surprising power of small language models. https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *Preprint*, arXiv:2310.06825.

Marcel A. Just and Patricia A. Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4):329.

Alan Kennedy, Robin Hill, and Joël Pynte. 2003. The Dundee corpus. In *Proceedings of the 12th European Conference on Eye Movement*.

Keren Gruteke Klein, Shachar Frenkel, Omer Shubi, and Yevgeni Berzak. 2025. Surprisal takes it all: Eye tracking based cognitive evaluation of text readability measures. *arXiv preprint arXiv:2502.11150*.

Reinhold Kliegl, Ellen Grabner, Martin Rolfs, and Ralf Engbert. 2004. Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, 16(1-2):262–284.

Victor Kuperman, Sascha Schroeder, and Daniil Gnetov. 2024. Word length and frequency effects on text reading are highly similar in 12 alphabetic languages. *Journal of Memory and Language*, 135:104497.

Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. 2017. lmertest package: tests in linear mixed effects models. *Journal of Statistical Software*, 82:1–26.

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

Steven G. Luke and Kiel Christianson. 2016. Limits on lexical prediction during reading. *Cognitive Psychology*, 88:22–60.

Steven G Luke and Kiel Christianson. 2018. The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 50:826–833.

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.

Maximilian M. Rabe, Dario Paape, Daniela Mertzen, Shravan Vasishth, and Ralf Engbert. 2024. Seam: An integrated activation-coupled model of sentence processing and eye movements in reading. *Journal of Memory and Language*, 135:104496.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners.

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

Keith Rayner. 2009. The 35th Sir Frederick Bartlett Lecture: Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology*, 62(8):1457–1506.

Keith Rayner and Charles Clifton. 2009. Language processing in reading and speech perception is fast and incremental: Implications for event-related potential research. *Biological Psychology*, 80(1):4–9.

Erik D. Reichle, Keith Rayner, and Alexander Pollatsek. 2003. The E-Z reader model of eye-movement control in reading: Comparisons to other models. *The Behavioral and Brain Sciences*, 26:445–526.

Brian D. Ripley. 2002. *Modern applied statistics with S*. Springer, New York, NY.

Lavinia Salicchi, Emmanuele Chersoni, and Alessandro Lenci. 2023. A study on surprisal and semantic relatedness for eye-tracking data prediction. *Frontiers in Psychology*, 14:1112365.

F. E. Satterthwaite. 1946. An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2(6):110–114.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

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.

Xinying Song, Alex Salcianu, Yang Song, Dave Dopson, and Denny Zhou. 2021. Fast WordPiece tokenization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2089–2103, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Débora Torres, Wagner R. Sena, Humberto A. Carmona, André A. Moreira, Hernán A. Makse, and José S Andrade Jr. 2021. Eye-tracking as a proxy for coherence and complexity of texts. *PlOS One*, 16(12):e0260236.

Emiel Van den Hoven, Franziska Hartung, Michael Burke, and Roel M. Willems. 2016. Individual differences in sensitivity to style during literary reading: Insights from eye-tracking. *Collabra*, 2(1):25.

Aaron Veldre, Lili Yu, Sally Andrews, and Erik D. Reichle. 2020. Towards a complete model of reading: Simulating lexical decision, word naming, and sentence reading with Über-Reader. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*. Cognitive Science Society.

Saranya Venkatraman, He He, and David Reitter. 2023. How do decoding algorithms distribute information in dialogue responses? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 953–962, Dubrovnik, Croatia. Association for Computational Linguistics.

Abraham Wald. 1943. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 54(3):426–482.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2024. WizardLM: Empowering large language models to follow complex instructions. In *12th International Conference on Learning Representations (ICLR 2024)*.

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, Singapore. Association for Computational Linguistics.

Appendix for Genre Matters: How Text Types Interact with Decoding Strategies and Lexical Predictors in Shaping Reading Behavior

A Log-Transformation of Reading Times

One of the core assumptions of linear models and linear-mixed models is that the residuals are normally distributed. This assumption is typically violated in eye-tracking data when raw reading times are used due to their right-skewed distribution. To corroborate the necessity of log-transforming the continuous reading measures, we conduct a Box-Cox transformation analysis (Ripley, 2002) to determine the most appropriate transformation for our continuous dependent variables. Specifically, we fit a linear model f_θ defined as

$$f_\theta := y_{ij} \sim \theta_0 + \theta_1 l_i + \theta_2 f_i + \theta_3 s_i + \theta_4 tt_i, \tag{2}$$

where y_{ij} are the raw (*i.e.*, non-log-transformed) total fixation times (TFTs) of word i read by subject j, l_i, f_i, and s_i are the word length, lexical frequency, and surprisal of word i, tt_i is the text type of the text to which word i belongs, θ_0 is the intercept, and $\theta_1, \ldots, \theta_4$ are the coefficients. We then apply the `boxcox()` function from the `MASS` library (Ripley, 2002) over lambda values ranging from -2 to 2 to estimate the optimal transformation parameters. This transformation is used to identify an optimal power transformation to stabilize the variance and make the residuals of a linear model more normally distributed. The resulting Box-Cox plot is depicted in Figure 4, where the x-axis represents different values of λ and the y-axis shows the log-likelihood of the model under each corresponding λ. The resulting Box-Cox profile depicts a peak near $\lambda = 0$ as well as a relatively smooth and parabolic curve, indicating that a logarithmic transformation of the response variable may best stabilize variance and improve model normality.

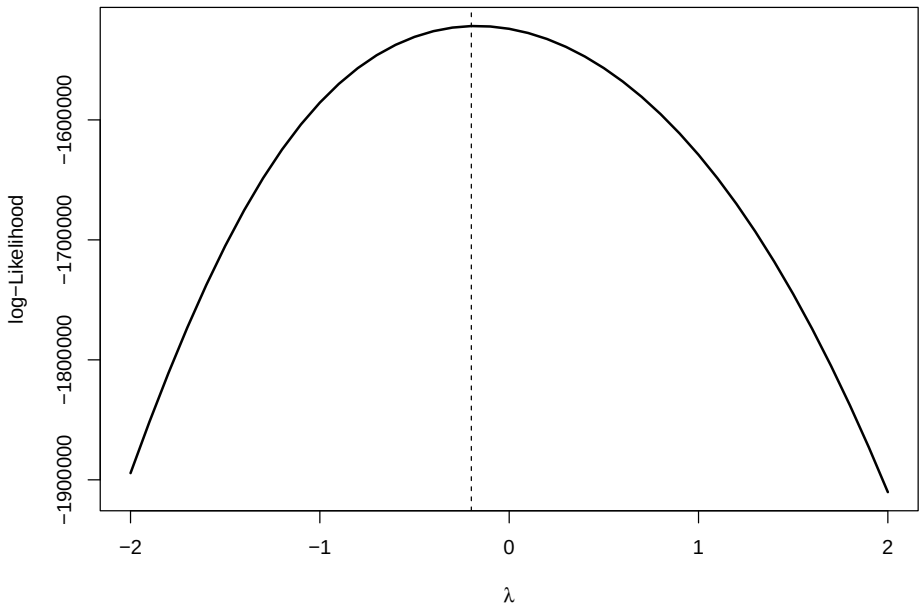

Figure 4: Box-Cox transformation analysis for the reading time variable TFT. The curve shows the profile log-likelihood of the linear model as a function of the transformation parameter λ.

B Statistical Significance Tests

In this study, we used the `lmer()` and `glmer()` functions from the `lme4` package (Bates et al., 2015) to fit linear and generalied linear mixed-effects models, respectively, depending on whether the dependent variable was continuous or binary. For models fit with `lmer()`, we used the `lmerTest` package (Kuznetsova et al., 2017), which provides p-values for fixed effects based on the Satterthwaite's approximation (Satterthwaite, 1946) of degrees of freedom. This method estimates degrees of freedom based on the structure of the random effects and accounts for the uncertainty introduced by them, yielding more accurate inferential statistics in mixed models. The significance of each fixed effect was then assessed using a two-sided t-test, where the test statistic is compared against a t-distribution with approximated degrees of freedom. For models fit with `glmer()`, the p-values are based on Wald z-tests (Wald, 1943), which assume asymptotic normality of the estimates.

C Pooling of Surprisal

The word-level surprisal values utilized in this study are already contained within EMTeC (Bolliger et al., 2025), where surprisal has been estimated with a range of language models, including GPT-2 *small* (Radford et al., 2019). Since language models employ tokenizers that separate words into sub-word tokens (Sennrich et al., 2016; Song et al., 2021) but the reading measure data is on word-level, the surprisal values must be pooled to word level.

Since the sum of two logarithms is equal to the logarithm of the product of their arguments, *i.e.*, $\log a + \log b = \log [a \cdot b]$, surprisal is pooled to word-level as follows: given k sub-word tokens $w_n, w_{n+1}, \ldots, w_{n+k}$ that belong to the same word token w, the word-level surprisal of w is computed as

$$\begin{aligned} s(w_n, w_{n+1}, \ldots, w_{n+k}) &= -\log p(w_n, w_{n+1}, \ldots, w_{n+k} \mid \boldsymbol{w}_{<n}) \\ &= -\log \Big[p(w_n \mid \boldsymbol{w}_{<n}) \cdot p(w_{n+1} \mid \boldsymbol{w}_{<n+1}) \cdot \ldots \cdot p(w_{n+k} \mid \boldsymbol{w}_{<n+k}) \Big] \\ &= -\log p(w_n \mid \boldsymbol{w}_{<n}) - \log p(w_{n+1} \mid \boldsymbol{w}_{<n+1}) \\ &\quad - \ldots - \log p(w_{n+k} \mid \boldsymbol{w}_{<n+k}). \end{aligned}$$

This shows that summing up sub-word level surprisal values is equivalent to computing the surprisal of the joint probability distribution of the sub-word tokens.

D Contrast Coding and Contrast Matrices

Contrast coding is a statistical technique used to analyze categorical factors (variables with discrete levels) in linear regression models. When we include categorical factors in regression analyses, we cannot directly use text labels like *fiction* or *poetry* as predictors because regression models require numerical input. Instead, we must convert these categories into numerical variables through contrast coding. This process transforms a factor with k levels into $k - 1$ numerical predictor variables that can be included in the regression model. Each contrast represents a specific comparison or hypothesis about differences between factor levels.

The choice of contrast coding scheme determines how we interpret the regression coefficients and what specific hypotheses we test. Different coding schemes answer different research questions: treatment contrasts compare each level to a baseline condition, repeated contrasts compare neighboring levels in sequence, and sum contrasts compare each level to the overall average. The coding scheme directly affects the meaning of the intercept term and the interpretation of main effects, particularly when interactions between factors are present in the model.

Sum contrast coding is one specific type of contrast coding that compares each factor level to the grand mean (average) across all levels of that factor. In sum contrasts, each of the $k - 1$ contrast variables tests whether a particular factor level differs significantly from the overall average performance across all conditions. For example, with the `text type` factor containing six levels—*non-fiction*, *fiction*, *poetry*, *summarization*, *article*, and *key-word text*—sum contrasts would create five contrast variables. Each would test whether a specific text type (e.g., fiction reading scores, poetry reading scores, non-fiction reading

scores, etc.) differs from the average reading score across all six text types. The sixth level (*key-word text*) serves as the reference category and is implicitly represented in the intercept term, which estimates the grand mean across all text types.

The mathematical implementation of sum contrasts uses specific coefficient patterns: the level being compared receives a coefficient of $+1$, while all other levels receive coefficients of $-\frac{1}{k}$ (where k is the total number of levels), ensuring that the contrast coefficients sum to zero, which is a requirement for centered contrasts. This coding scheme allows us to interpret regression coefficients as deviations from the grand mean, making the results particularly intuitive for factorial designs where we want to understand how each experimental condition performs relative to the overall average performance.

Below the contrast matrices used in the experiments are depicted. Table 1 shows the sum-contrast coded factor `text type`, and Table 2 shows the sum-contrast coded factor `decoding strategy`.

Table 1: Sum contrast matrix for the factor `text type`.

Factor Level	Article synopsis vs grand-mean	Summarization vs grand-mean	Non-fiction vs grand-mean	Fiction vs grand-mean	Poetry vs grand-mean
Article synopsis	1	0	0	0	0
Summarization	0	1	0	0	0
Non-fiction	0	0	1	0	0
Fiction	0	0	0	1	0
Poetry	0	0	0	0	1
Key-word text	-1	-1	-1	-1	-1

Table 2: Sum contrast matrix for the factor `decoding strategy`.

Factor Level	Beam search vs grand mean	Sampling vs grand mean	Top-k vs grand mean	Top-p vs grand mean
Beam search	1	0	0	0
Sampling	0	1	0	0
Top-k	0	0	1	0
Top-p	0	0	0	1
Greedy search	-1	-1	-1	-1

E RQ_2

The fixed effects of the psycholinguistic predictors are plotted in Figure 5 as a sanity check. Across all predictors and reading measures, the direction of the effect is as expected: the effects of lexical frequency are significantly negative (high-frequency words cause lower reading times), the effects of surprisal are significantly positive (high-surprisal words cause longer reading times), as are the effects of word length (longer words cause longer reading times). The only exception is surprisal as a predictor for the binary variable first-pass regression.

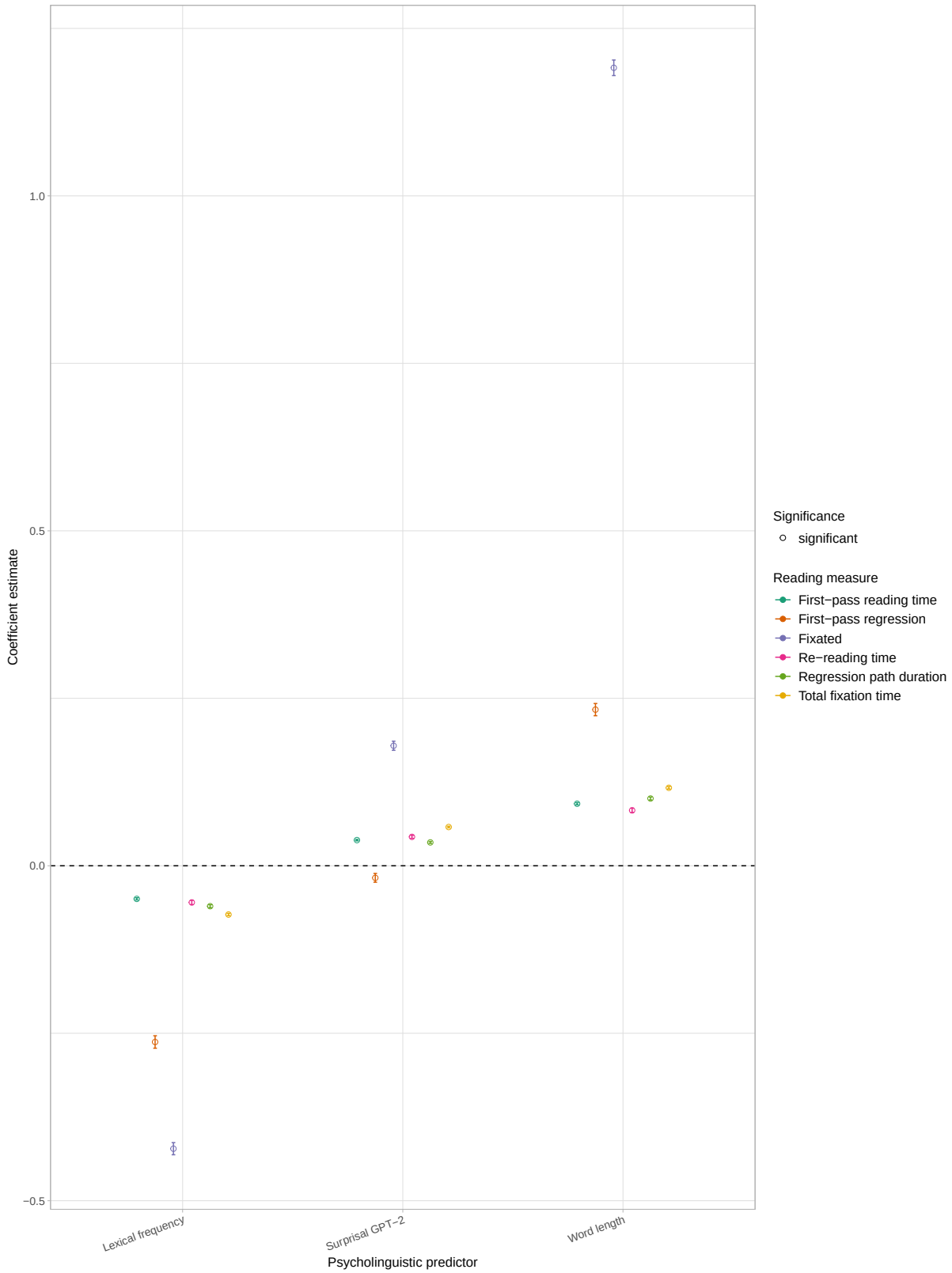

Figure 5: Estimates (mean and 95% CI) of the fixed effects of the psycholinguistic predictors lexical frequency, word length, and surprisal. All effects are significantly different from zero.

F RQ_3

Figure 6 depicts the estimates of the fixed effects of the psycholinguistic predictors and the sum-contrast coded predictors text type and decoding strategy. This serves as a sanity check to corroborate that the effects of the psycholinguistic predictors are as would be expected: the effects of lexical frequency are negative (frequent words cause lower reading times), the effects of word length are positive (longer words cause longer reading times), as are the effects of surprisal (high-surprisal words cause longer reading times). Moreover, the main effects of the text types exhibit the same pattern as in the results for RQ_1 (see § 3.4). The main effects of the different decoding strategies, on the other hand, are mostly not significantly different from the grand mean with the exception of beam search, indicating that texts generated with this decoding strategy elicit longer-than-average re-reading time.

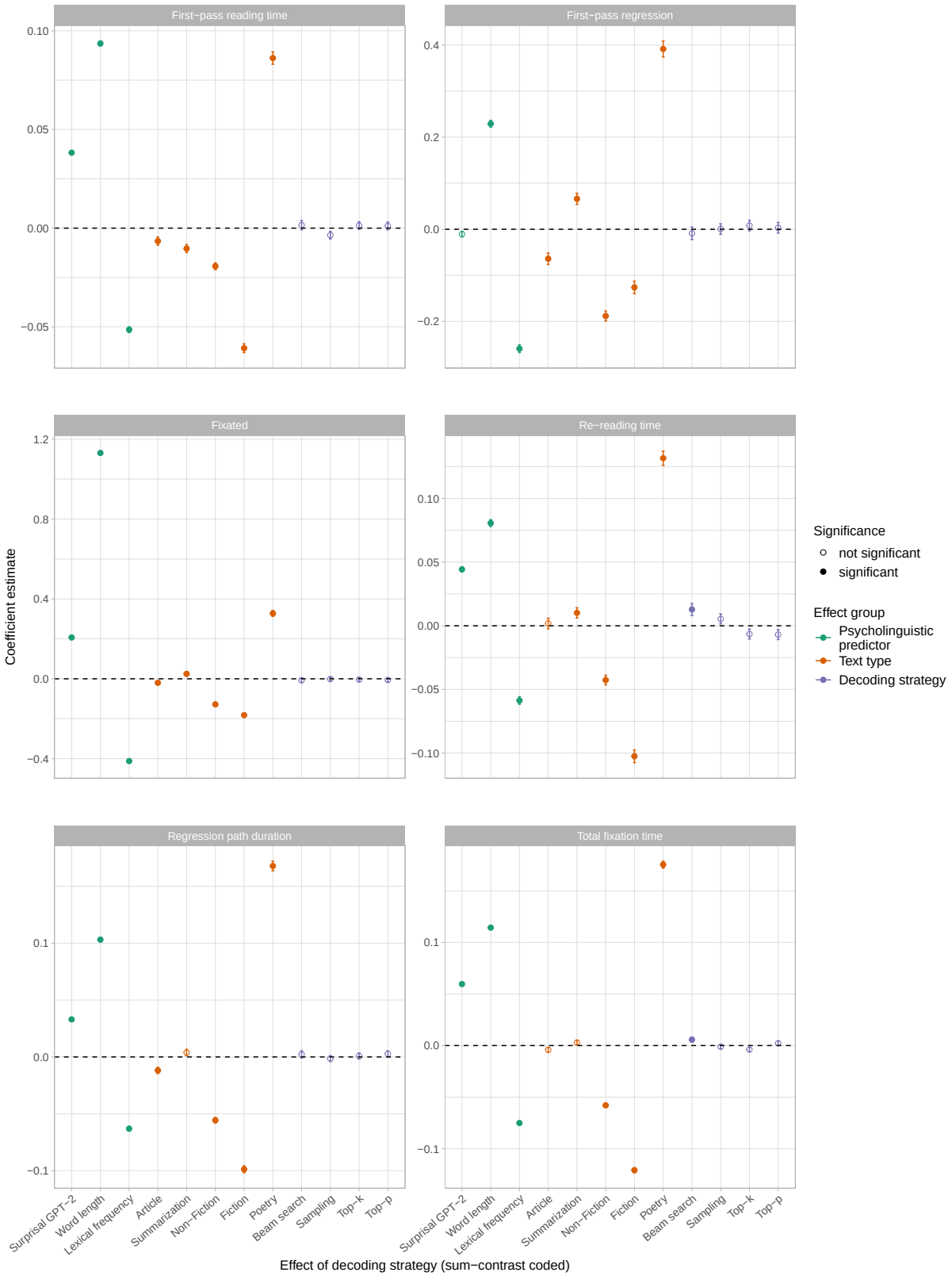

Figure 6: Estimates (mean and 95% CI) of the fixed effects of the psycholinguistic predictors lexical frequency, word length, and surprisal, and of the sum-contrast coded factors text type and decoding strategies. A filled dot indicates that the effect is significantly different from zero for the continuous psycholinguistic predictors, or significantly different from the grand mean for the sum-contrast coded text type and decoding strategy.