On language models' cognitive biases in reading time prediction

Patrick Haller [1] **Lena S. Bolliger** [1] **Lena A. Jäger** [1 2]

Abstract

To date, most investigations on surprisal and entropy effects in reading have been conducted on the group-level, disregarding individual differences. In this work, we revisit the predictive power (PP) of different language models' (LMs') surprisal and entropy measures on data of human reading times by incorporating information of language users' cognitive capacities. To do so, we assess the PP of surprisal and entropy estimated from generative LMs on reading data from subjects for which scores from psychometric tests targeting different cognitive domains are available. Specifically, we investigate if modulating surprisal and entropy relative to the readers' cognitive scores increases prediction accuracy of reading times, and we examine whether LMs exhibit systematic biases in the prediction of reading times for cognitively high- or low-scoring groups, allowing us to investigate what type of psycholinguistic subjects a given LM emulates. We find that incorporating cognitive capacities mostly increases PP of surprisal and entropy on reading times, and that individuals performing high in cognitive tests are less sensitive to predictability effects. Our results further suggest that the analyzed LMs emulate readers with lower verbal intelligence, suggesting that for a given target group (i.e., individuals with high verbal intelligence), these LMs provide less accurate predictability estimates. Finally, our study underlines the value of incorporating individual-level information to gain insights into how LMs operate internally.

1 Introduction

Human language comprehension and, by extension, human reading is incremental in nature: humans process words sequentially (Rayner & Clifton Jr, 2009), and different words in varying contexts impose different amounts of cognitive processing efforts. Similarly, language models' conditional probability distributions assign different probabilities for potential continuations for a given prefix. The relationship between cognitive effort and predictability measures derived from LMs was operationalized in *surprisal theory* (Hale, 2001; Levy, 2008). Since then, a large body of research has investigated the details of the relationship between surprisal and entropy, and human processing effort (Linzen & Jaeger, 2016; Kuribayashi et al., 2021; de Varda & Marelli, 2022; Wilcox et al., 2023b; Shain et al., 2024, i.a.). So far, most studies have tested these predictions on the group-level, neglecting individual cognitive differences that might influence readers' capacities to make predictions about upcoming material.

In this work, we revisit the relationship between both surprisal and contextual entropy and data of human processing effort by considering language users' individual cognitive differences. More specifically, we examine whether cognitive capacities induce a higher surprisal or entropy effect for individuals with a certain cognitive profile, whether the predicted individual effect is similar across different language models (LMs), and the impact of cognitive scores on the predictive power of surprisal and entropy estimated from a range of LMs. We therefore investigate the following hypotheses:

- $\mathbf{H}_1$: Modulating surprisal and entropy effects relative to individual cognitive capacities improves the predictive power on reading times on unseen data.
- $\mathbf{H}_2$: Individuals with higher cognitive performance exhibit a lower surprisal or entropy effect.
- $\mathbf{H}_3$: LMs are significantly better at predicting reading times for certain cognitive profiles.

To address these hypotheses, we utilize the *Individual Differences Corpus* (InDiCo; Haller et al., 2023), which contains both reading data and scores of a comprehensive psychometric assessment targeting various cognitive capacities, including: verbal and non-verbal working memory, verbal and non-verbal cognitive control, verbal and non-verbal intelligence, and reading fluency. We deploy five pre-trained generative LMs from three language-families—GPT2 base and large, Llama 7B and 13B, and Mixtral—to estimate both surprisal and contextual entropy and quantify their predictive power by including them as predictors in linear

[1]Department of Computational Linguistics, University of Zurich, Switzerland [2]Department of Computer Science, University of Potsdam, Germany. Correspondence to: Patrick Haller <haller@cl.uzh.ch>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

regressors, fitted to predict by-word reading times from InDiCo. We then assess the regressors' log-likelihood after including these predictors and their interaction with the psychometric scores against a baseline model.

2 Related work

2.1 Predictive power of surprisal and entropy

Surprisal (Hale, 2001; Levy, 2008) is a measure of predictability of a word and shown to be proportional to cognitive effort in human sentence processing. It is quantified as the negative log probability of a word given its preceding context. Since the formalization of surprisal theory, many studies have corroborated that surprisal correlates with reading times (Demberg & Keller, 2008; Shain, 2021; Hoover et al., 2023; Pimentel et al., 2023) and deploying different LMs (Wilcox et al., 2020; 2023a; Goodkind & Bicknell, 2018). Recently, Oh & Schuler (2023b) revealed that large models, despite their lower perplexity, provide worse PP of RTs, and Oh & Schuler (2023a) further demonstrated that LMs provide the best fit to RTs when trained on around 2B tokens. Linzen & Jaeger (2016) first examined how sentence processing is affected by readers' uncertainty, measured via entropy or entropy reduction, and found that contextual entropy does not correlate with reading times. Later, van Schijndel & Schuler (2017) showed that entropy is indeed predictive of RTs. Wilcox et al. (2023b) later found that adding entropy as an additional predictor (while keeping surprisal) improves the model's PP, while replacing surprisal with entropy leads to a decrease in PP. However, Pimentel et al. (2023) also show that using contextual entropy as a predictor can be as good as surprisal when analyzing *anticipatory* effects reflected in skipping rates, as opposed to *responsive* effects captured by gaze duration.

2.2 Individual differences in sentence processing

Theories of sentence processing generally assume that the cognitive mechanisms involved in language processing are qualitatively identical across speakers. However, this perspective has been challenged, and evidence is emerging that differences in cognitive abilities among language users do indeed have a significant impact on processing (Vuong & Martin, 2014; Nicenboim et al., 2015; Farmer et al., 2017, i.a.). For instance, Kuperman & Van Dyke (2011) demonstrate that measures related to cognitive control interact with word length and lexical frequency effects on fixation times, and Nicenboim et al. (2015) show that readers ranking lower in working-memory tests exhibit more regressive saccades in regions with high memory load.

Several studies have also investigated individual differences in surprisal effects, particularly in the realm of native and non-native reading (Berzak & Levy, 2023; Schneider et al., 2023). For instance, Berzak & Levy (2023) demonstrate that higher L2 proficiency is associated with increased sensitivity to a word's predictability in context (surprisal). Moreover, Škrjanec et al. (2023) show that specialized surprisal from domain-adapted LMs improves reading-time predictions for expert readers.

3 Methods

3.1 Estimating predictability effects

Given a vocabulary Σ and an augmented vocabulary $\bar{\Sigma} = \Sigma \cup \{\text{EOS}\}$, which contains a special EOS (end-of-sentence) token, the **surprisal** (Shannon, 1948) of a given sequence is defined as

$$s(u_n) \stackrel{\text{def}}{=} -\log p(u_n \mid \mathbf{u}_{<n}), \tag{1}$$

where $p(\cdot \mid \mathbf{u}_{<n})$ is the true distribution over words $u \in \bar{\Sigma}$ in context $\mathbf{u}_{<n}$. Since we do not have access to the true distribution $p(\cdot \mid \mathbf{u}_{<n})$, we approximate it using an autoregressive LM.

The **contextual entropy** of a $\bar{\Sigma}$-valued random variable U_n at index n is the expected value of its surprisal, i.e.:

$$\begin{aligned} \mathrm{H}(U_n \mid \mathbf{U}_{<n} = \mathbf{u}_{<n}) &\stackrel{\text{def}}{=} \mathbb{E}_{u \sim p(\cdot \mid \mathbf{u}_{<n})}\left[s_n(u)\right] \\ &= -\sum_{u \in \bar{\Sigma}} p(u|\mathbf{u}_{<n}) \log_2 p(u|\mathbf{u}_{<n}). \end{aligned} \tag{2}$$

It is a specific version of the Shannon entropy $\mathrm{H}(U) \stackrel{\text{def}}{=} -\sum_{u \in \mathcal{U}} p(u) \log p(u)$, conditioned on the left context.

3.2 Assessing predictive power

We utilize linear-mixed models (LMMs) $\mathcal{M}$ to predict a reading time measure y_{ij}, obtained from a subject j on word i, from a set of standardized word-level and subject-level predictors $\mathbf{x}_{ij}$, *i.e.,* $\mathcal{M} : \mathbf{x}_{ij} \mapsto y_{ij}$.

For our analyses, we aim to quantify the predictive power of a given predictor of interest x^q (*e.g.,* surprisal). To do so, we first define a baseline model $\mathcal{M}^b : \mathbf{x}^b_{ij} \mapsto y_{ij}$ that includes a set of baseline predictors $\mathbf{x}^b_{ij}$, and a target model $\mathcal{M}^t : \mathbf{x}^b_{ij} \oplus x^q_{ij} \mapsto y_{ij}$ that additionally includes the predictor of interest x^q_{ij}, where $\oplus$ represents the concatenation of two sets of predictors. Following previous work (Wilcox et al., 2020, i.a.), we operationalize the predictive power as the mean difference in log-likelihood (Δ_{LL}) between the target and the baseline model. To avoid overfitting, we perform 10-fold cross validation. A positive Δ_{LL} indicates a better fit of the target model to the data.

4 Experiments

Data. We employ German eye-tracking-while-reading data from InDiCo (Haller et al., 2023). In addition to the reading data from 61 native German speakers, the corpus contains a battery of individual psychometric scores in four cognitive domains: cognitive control, working memory, intelligence, and reading fluency. Specifically, we use the *first-pass reading times* (FPRT), as well as the standardized scores of 13 psychometric tests. For a detailed description of the data, see Appendix B.

	Cognitive domain	Test	Effect size of interaction term GPT-2 *base*	GPT-2 *large*	Llama-2 7b	Llama-2 13b	Mixtral
Entropy	Cognitive control	FAIR	$-0.002\,(\pm 0.001)^{\dagger}$	$-0.001\,(\pm 0.001)^{\dagger}$	$-0.003\,(\pm 0.001)^{\dagger}$	$-0.003\,(\pm 0.001)^{\dagger}$	$-0.003\,(\pm 0.001)^{\dagger}$
		Simon	$0.003\,(\pm 0.001)^{\dagger}$	$0.002\,(\pm 0.001)^{\dagger}$	$0.005\,(\pm 0.001)^{\dagger}$	$0.004\,(\pm 0.001)^{\dagger}$	$0.005\,(\pm 0.001)^{\dagger}$
		Stroop	$-0.001\,(\pm 0.001)^{\dagger}$	$-0.001\,(\pm 0.001)^{\dagger}$	$0\,(\pm 0.001)$	$0\,(\pm 0.001)$	$0\,(\pm 0.001)$
	Intelligence	MWT	$-0.006\,(\pm 0.001)$	$-0.005\,(\pm 0.001)^{\dagger}$	$-0.008\,(\pm 0.001)$	$-0.008\,(\pm 0.001)$	$-0.009\,(\pm 0.001)$
		RIAS non-verbal	$0\,(\pm 0.001)^{\dagger}$	$0\,(\pm 0.001)$	$0\,(\pm 0.001)$	$0\,(\pm 0.001)$	$-0.001\,(\pm 0.001)^{\dagger}$
		RIAS total	$-0.005\,(\pm 0.001)^{\dagger}$	$-0.004\,(\pm 0.001)^{\dagger}$	$-0.005\,(\pm 0.001)^{\dagger}$	$-0.005\,(\pm 0.001)^{\dagger}$	$-0.006\,(\pm 0.001)$
		RIAS verbal	$-0.007\,(\pm 0.001)$	$-0.005\,(\pm 0.001)$	$-0.007\,(\pm 0.001)$	$-0.007\,(\pm 0.001)$	$-0.007\,(\pm 0.001)$
	Reading fluency	SLRT pseudo-words	$-0.006\,(\pm 0.001)$	$-0.004\,(\pm 0.001)^{\dagger}$	$-0.008\,(\pm 0.001)$	$-0.007\,(\pm 0.001)$	$-0.007\,(\pm 0.001)$
		SLRT words	$-0.005\,(\pm 0.001)^{\dagger}$	$-0.003\,(\pm 0.001)^{\dagger}$	$-0.009\,(\pm 0.001)$	$-0.007\,(\pm 0.001)$	$-0.007\,(\pm 0.001)$
	Working memory	Memory updating	$-0.003\,(\pm 0.001)^{\dagger}$	$-0.002\,(\pm 0.001)^{\dagger}$	$-0.004\,(\pm 0.001)^{\dagger}$	$-0.003\,(\pm 0.001)^{\dagger}$	$-0.003\,(\pm 0.001)^{\dagger}$
		Operation span	$-0.005\,(\pm 0.001)^{\dagger}$	$-0.003\,(\pm 0.001)^{\dagger}$	$-0.008\,(\pm 0.001)$	$-0.007\,(\pm 0.001)$	$-0.008\,(\pm 0.001)$
		Sentence span	$-0.003\,(\pm 0.001)^{\dagger}$	$-0.002\,(\pm 0.001)^{\dagger}$	$-0.007\,(\pm 0.001)$	$-0.006\,(\pm 0.001)$	$-0.007\,(\pm 0.001)$
		Spatial short-term memory	$-0.001\,(\pm 0.001)^{\dagger}$	$0\,(\pm 0.001)^{\dagger}$	$0.002\,(\pm 0.001)^{\dagger}$	$0.001\,(\pm 0.001)^{\dagger}$	$0\,(\pm 0.001)^{\dagger}$
Surprisal	Cognitive control	FAIR	$-0.01\,(\pm 0.001)$	$-0.009\,(\pm 0.001)$	$-0.008\,(\pm 0.001)$	$-0.008\,(\pm 0.001)$	$-0.007\,(\pm 0.001)$
		Simon	$0.01\,(\pm 0.001)$	$0.01\,(\pm 0.001)$	$0.009\,(\pm 0.001)$	$0.008\,(\pm 0.001)$	$0.007\,(\pm 0.001)$
		Stroop	$-0.001\,(\pm 0.001)^{\dagger}$	$-0.001\,(\pm 0.001)^{\dagger}$	$-0.001\,(\pm 0.001)^{\dagger}$	$0\,(\pm 0.001)$	$0\,(\pm 0.001)$
	Intelligence	MWT	$-0.016\,(\pm 0.001)$	$-0.015\,(\pm 0.001)$	$-0.014\,(\pm 0.001)$	$-0.014\,(\pm 0.001)$	$-0.014\,(\pm 0.001)$
		RIAS non-verbal	$0\,(\pm 0.001)^{\dagger}$	$0\,(\pm 0.001)$	$0.001\,(\pm 0.001)^{\dagger}$	$0\,(\pm 0.001)$	$0.001\,(\pm 0.001)^{\dagger}$
		RIAS total	$-0.011\,(\pm 0.001)$	$-0.011\,(\pm 0.001)$	$-0.009\,(\pm 0.001)$	$-0.009\,(\pm 0.001)$	$-0.007\,(\pm 0.001)$
		RIAS verbal	$-0.015\,(\pm 0.001)$	$-0.014\,(\pm 0.001)$	$-0.012\,(\pm 0.001)$	$-0.012\,(\pm 0.001)$	$-0.01\,(\pm 0.001)$
	Reading fluency	SLRT pseudo-words	$-0.018\,(\pm 0.001)$	$-0.017\,(\pm 0.001)$	$-0.015\,(\pm 0.001)$	$-0.014\,(\pm 0.001)$	$-0.012\,(\pm 0.001)$
		SLRT words	$-0.019\,(\pm 0.001)$	$-0.017\,(\pm 0.001)$	$-0.015\,(\pm 0.001)$	$-0.014\,(\pm 0.001)$	$-0.012\,(\pm 0.001)$
	Working memory	Memory updating	$-0.011\,(\pm 0.001)$	$-0.01\,(\pm 0.001)$	$-0.008\,(\pm 0.001)$	$-0.008\,(\pm 0.001)$	$-0.006\,(\pm 0.001)$
		Operation span	$-0.018\,(\pm 0.001)$	$-0.018\,(\pm 0.001)$	$-0.015\,(\pm 0.001)$	$-0.015\,(\pm 0.001)$	$-0.012\,(\pm 0.001)$
		Sentence span	$-0.016\,(\pm 0.001)$	$-0.015\,(\pm 0.001)$	$-0.014\,(\pm 0.001)$	$-0.013\,(\pm 0.001)$	$-0.012\,(\pm 0.001)$
		Spatial short-term mem.	$0\,(\pm 0.001)^{\dagger}$	$0\,(\pm 0.001)$	$0.001\,(\pm 0.001)^{\dagger}$	$0.001\,(\pm 0.001)^{\dagger}$	$0.001\,(\pm 0.001)^{\dagger}$

Table 1: Effect sizes of interaction terms $\pm$ standard error between entropy(top)/surprisal(bottom) and psychometric test scores.† indicates that the inclusion of the interaction term did not lead to a significant increase or decrease in Δ_{LL}.

Word-level predictors. To extract surprisal and contextual entropy estimates, we deploy the German versions of five pretrained transformer-based LMs of different families and sizes, namely GPT-2 *base* and *large* (Radford et al., 2019), Llama 2 7b and 13b (Touvron et al., 2023), and Mixtral (Jiang et al., 2024). For details, see Appendix A.1. We compute word-level surprisal by summing the surprisal values of the sub-word tokens (Sennrich et al., 2016; Song et al., 2021). Similarly, to obtain the word-level contextual entropy, we use the sum of the sub-word token-level contextual entropy values as proxy for the joint entropy of the sub-word tokens' distributions (see Appendix A.2 for details).

We further include lexical frequency and word length in our analyses as they are known to have an impact on human reading behavior (see Appendix A.1 for details).

Psychometric scores. The psychometric assessment in IndiCo includes a total of 13 tests targeting different cognitive domains such as verbal and non-verbal working memory, cognitive control and intelligence, as well as reading fluency. A list of tests and their abbreviations can be found in Appendix B. We transform all test scores such that higher scores indicate higher performance.

4.1 Assessing the PP and magnitude of interactions between surprisal/entropy and psychometric scores ($\mathbf{H_{1,2}}$)

First, we investigate whether the interaction between cognitive scores and surprisal, or entropy, leads to an increase in predictive power on reading times (H_1). We define a baseline model $\mathcal{M}_1^b$ with predictors $\mathbf{x}_{ij}^{b_1}$ including the word-level predictors l_i (word length), f_i (lexical frequency), s_i, h_i, and the subject-level predictor c_j denoting the test score of a specific psychometric test (e.g., *word-reading fluency*) obtained for subject j. We additionally include a by-subject intercept β_{0j}, thus, $\mathcal{M}_1^b : y_{ij} \sim \beta_0 + \beta_{0j} + \beta_1\, l_i + \beta_2\, f_i + \beta_3\, s_i + \beta_4\, h_i + \beta_5\, c_j.$, where y_{ij} refers to the log-transformed first-pass reading time of subject j for the i^{th} word in the stimulus corpus across all texts.

The target models are defined as $\mathcal{M}_1^{t_s}$ and $\mathcal{M}_1^{t_h}$, including an additional interaction term between either surprisal or entropy and a given psychometric score c_j (e.g., *word-reading fluency score*) obtained for subject j, $x_{ij}^{q_1} \in \{s_i \cdot c_j, h_i \cdot c_j\}$: $\mathcal{M}_1^t : y_{ij} \sim \beta_0 + \beta_{0j} + \beta_1\, l_i + \beta_2\, f_i + \beta_3\, s_i + \beta_4\, h_i + \beta_5\, c_j + \beta_6\, x_{ij}^{q_1}$. A positive Δ_{LL} between the target and the baseline model indicates that including the participant's score of a given psychometric test improves the prediction on the held-out test data. We run paired permutation tests to establish whether a given Δ_{LL} is significantly different from 0 at $\alpha = .05$. We then re-run the target models $\mathcal{M}_1^{t_s}$ and $\mathcal{M}_1^{t_h}$ on the *entire* dataset to examine the effect sizes (coefficients) of the interaction term between the scores and the surprisal and entropy estimates, β_6.

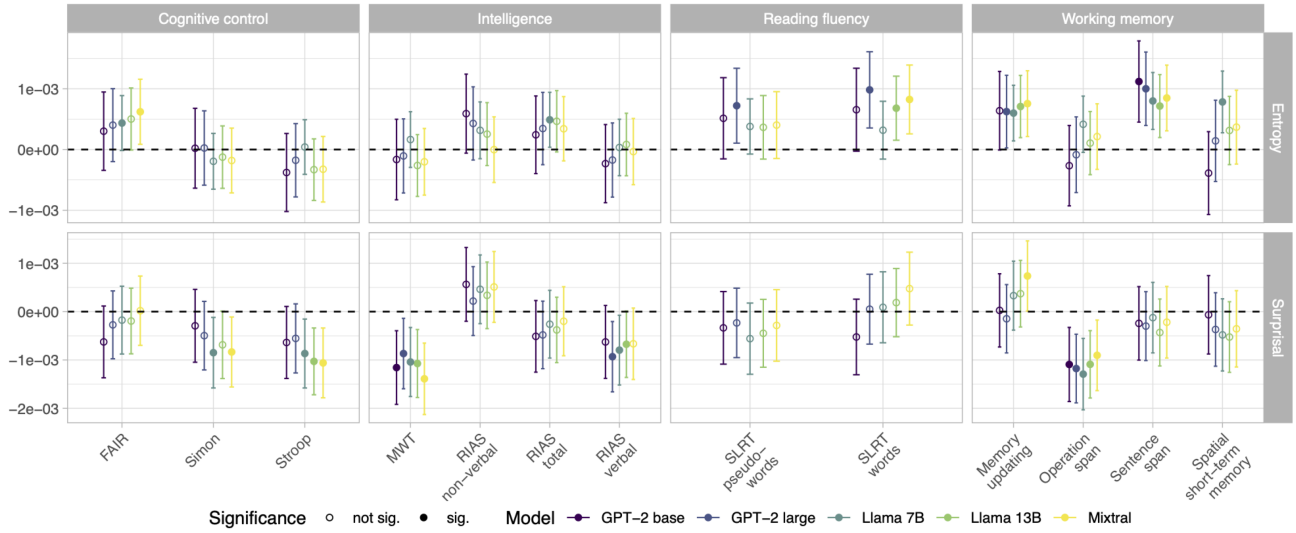

Figure 1: Difference in PP (ΔPP) (mean and 95% CI) of surprisal and contextual entropy for reading times. Positive ΔPP indicate higher PP for high-performing individuals; negative ΔPP indicates higher PP for low-performing individuals.

Results. We present the effect sizes of the interaction between scores and predictability measures in Tab.1.[1] Coloured cells indicate significant increases in PP. Overall, the interaction terms between surprisal/entropy and most psychometric scores lead to significant increases in PP, except for Stroop, non-verbal RIAS and spatial short-term memory. Notably, PP is not significant or, if significant, extremely small for these three scores across all models. Additionally, there are notable differences among different cognitive domains: modulating surprisal with scores targeting reading fluency or the working-memory span tests yields the highest predictive power, followed by verbal intelligence scores. Finally, we note that interactions with surprisal extracted from the GPT-2 family have the highest PP. Conversely, interactions with GPT-based entropy have the lowest PP.

Next, we assess the magnitude of the interaction term coefficients (H_2). We notice that for a given psychometric test, all models consistently modulate surprisal and entropy effects in the same direction. For most psychometric tests, higher scores result in a reduction of surprisal and entropy effects, indicated by the negative interaction term coefficients. This suggests that individuals with higher scores show lower sensitivity to a word's predictability. This holds true across all tests, the only exception being the Simon test, providing a measure of non-verbal inhibitory cognitive control. Here, high-performing individuals exhibit larger surprisal effects. Positive coefficients are also found for the Stroop task and the non-verbal part of the RIAS (intelligence), although they are extremely small.

[1]Fig. 4 additionally shows the Δ_{LL} across all psychometric tests and models.

4.2 Assessing the difference in predictive power between cognitive profiles (H_3)

Finally, we investigate whether there are differences in the predictive power of LM surprisal and entropy for reading times obtained from individuals with different cognitive profiles. In other words, we ask the question what type of psycholinguistic subject a given language model emulates. To do so, we split the reading time data into subsets of high-performing ($\uparrow$) and low-performing individuals ($\downarrow$) at the median of each score. Then, for each group, we compute the Δ_{LL} between the baseline model $\mathcal{M}_3^b$ and the target model $\mathcal{M}_3^t$ with an additional predictor of interest $x_i^{q_3} \in \{s_i, h_i\}$, i.e. either surprisal or entropy. The individual $\Delta_{\text{LL}\downarrow}$ and $\Delta_{\text{LL}\uparrow}$ indicate the predictive power of surprisal and entropy for each group separately. In order to answer which group exhibited a higher relative gain in PP, we assess the difference in predictive power $\Delta\text{PP} \stackrel{\text{def}}{=} \Delta_{\text{LL}\uparrow} - \Delta_{\text{LL}\downarrow}$.

Results. Fig.1 presents the differences (mean and 95% CI) in predictive power (ΔPP) of surprisal or entropy between two groups that performed above or below the median, respectively, in a given psychometric test. $\Delta\text{PP} > 0$ indicates higher PP for the high-performing group, $\Delta\text{PP} < 0$ indicates higher PP for the low-performing group.

First, looking at the results for entropy, we note that across all models, entropy predicts the RTs of individuals among the high-performing groups in the memory-updating and operation-span tests significantly better. For surprisal, we find that across all models, RT predictions are significantly better for the low-performing group in the operation span test as well as the vocabulary size test MWT. Moreover, surprisal extracted from GPT-large and Llama 7B leads to significant gains in PP for the low-performing group in the RIAS test, which like MWT assesses verbal intelligence.

5 Discussion

In summary, our findings suggest that (1) individuals exhibit surprisal and entropy effects relative to certain cognitive capacities, and that (2) a given language model may have higher predictive power of reading times for individuals with a certain cognitive profile.

5.1 Implications for the cognitive mechanisms of language processing

In our first two experiments, we found a negative coefficient for the interaction terms between surprisal and reading fluency. Compared to other psychometric scores, these coefficients are relatively large (see Tab. 1) and the interaction terms' Δ_{LL} are high (see Fig. 4). The coefficient can be interpreted from two perspectives. From the participants' perspective, it underlines that individuals with high reading fluency exhibit lower surprisal effects. These results might indicate that less fluent readers rely more on predictive processing, hence their reading is easily interrupted by less predictable continuations, leading to longer reading times. Experienced readers, on the other hand, might be more trained to integrate unexpected material effortlessly. From the models' perspective, on the other hand, it means that LMs overestimate the surprisal effect exhibited by highly fluent readers. Similar arguments can be made for the verbal intelligence test (RIAS-verbal), which is correlated with reading fluency (cf. Figure 2).

Regarding working memory, the span tests (operation and sentence span) lead to substantial increases in PP, and the magnitude of their interaction terms indicate that individuals with higher scores in both tests show weaker surprisal effects. This might be explained by the fact that high working memory can be associated with the capability to hold competing continuations in memory, including less likely ones that sometimes turn out to be the actual continuation.

5.2 Cognitive profiles of language models

Regarding the bias analyses, the results presented in Figure 1 revealed that surprisal estimates across all tested models predicted RTs better for the group of individuals with low verbal intelligence scores, measured with two largely complementary tests: one that assesses word knowledge (MWT-B), and one that assesses verbal logical thinking via question answering and sentence completion (RIAS-verbal). At first glance, this result is surprising since a language model has been exposed to billions of tokens, and therefore, one might expect that it emulates a psycholinguistic subject with high verbal intelligence. However, a language model's predictions are always relative, i.e., even if it has seen infrequent words, it will still have a preference in terms of likelihood for the more regular, frequent continuation. Individuals with high verbal intelligence do not struggle with such contexts since they are very familiar even with uncommon terminology.

Additionally, we found that the PP of entropy is significantly higher for individuals with high working memory capacities, measured via memory updating and sentence span. This result suggests that uncertainty measures about upcoming material exhibited by LMs are more in line with the way high-working memory individuals process language, potentially driven by taking into account longer contexts, or keeping track of relevant long dependencies.

Even though most results from all three experiments are consistent within and across different LM families, there are exceptions. For instance, entropy estimated from GPT-2 large showed the strongest increase in PP for the high reading-fluency *word reading* group (Figure 1). For the high reading-fluency *pseudo-word reading* group, it represents the only measure with a significant increase in PP. This suggests that entropy extracted from GPT-2 large is a better proxy of processing effort for readers with lower verbal intelligence than entropy estimated with GPT-2 base. This illustrates that the choice of LM to estimate predictability measures is crucial for downstream analyses in psycholinguistic studies or NLP applications, especially when working with specific target groups. In such settings, it might be worthwhile considering a model that is less biased, or, in other words, whose predictability measures are well-aligned with the target group at hand as it will most likely lead to more accurate results.

While this study was aimed at uncovering model-internal biases, it might be worthwhile to, in turn, extend the investigation to whether text *produced* by a given LM is biased towards being processed more easily by individuals with specific cognitive characteristics. This is particularly important for tasks such as text summarization or simplification that might need to be tailored to specific groups.

6 Conclusion

To date, most investigations on predictability effects have been conducted on the group-level, assuming that the predictive power of next-word predictability metrics such as surprisal or entropy on human reading times is uniform across cognitive profiles. Our work illustrates how the use of LMs in the context of psycholinguistic studies can be reveal aspects about how these systems process language.

Acknowledgements

This work was partially funded by the Swiss National Science Foundation under grant 100015L_212276/1 (MeRID). We thank David Reich, Nora Hollenstein and Omer Shubi for valuable discussions regarding this work.

Impact Statement

Our work underscores the importance of considering individual-level information to better understand how LMs function internally and enhance their predictive accuracy

in modeling human reading behavior. As mentioned in the discussion, it needs to be investigated whether our findings are corroborated when studying whether text *produced* by a given LM is biased towards being processed differently by individuals with specific cognitive characteristics. Nevertheless, by understanding how individual cognitive differences influence reading comprehension and processing, educational tools can be tailored (i.e., by using specific LMs) to meet the specific needs of different learners, possibly leading to more effective teaching strategies and improved learning outcomes, particularly for students with diverse cognitive profiles.

From a Human-Computer Interaction point-of-view, incorporating individual differences into LMs could lead to more personalized and user-friendly interfaces. This can enhance user experience across various applications, and can eventually help in developing more fair and unbiased systems.

References

Berlin-Brandenburgische Akademie der Wissenschaften. DWDS – Digitales Wörterbuch der deutschen Sprache. Das Wortauskunftssystem zur deutschen Sprache in Geschichte und Gegenwart. http://www.dwds.de, 2016.

Berzak, Y. and Levy, R. Eye movement traces of linguistic knowledge in native and non-native reading. *Open Mind*, pp. 1–18, 2023.

de Varda, A. and Marelli, M. The effects of surprisal across languages: Results from native and non-native reading. In He, Y., Ji, H., Li, S., Liu, Y., and Chang, C.-H. (eds.), *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pp. 138–144, Online only, November 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-aacl.13.

Demberg, V. and Keller, F. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210, 2008.

Farmer, T. A., Fine, A. B., Misyak, J. B., and Christiansen, M. H. Reading span task performance, linguistic experience, and the processing of unexpected syntactic events. *Quarterly Journal of Experimental Psychology*, 70(3): 413–433, 2017.

Goodkind, A. and Bicknell, K. 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)*, pp. 10–18, 2018.

Hale, J. A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the North American Chapter of the Association for Computational Linguistics*, 2001.

Haller, P., Koncic, I., Reich, D., and Jäger, L. A. Measurement reliability of individual differences in sentence processing: A cross-methodological reading corpus and bayesian analysis. *ArXiv Preprint*, 2023.

Heister, J., Würzner, K.-M., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, A., and Kliegl, R. dlexdb–eine lexikalische datenbank für die psychologische und linguistische forschung. *Psychologische Rundschau*, 2011.

Hoover, J. L., Sonderegger, M., Piantadosi, S. T., and O'Donnell, T. J. The plausibility of sampling as an algorithmic theory of sentence processing. *Open Mind*, 7: 350–391, 2023.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., Casas, D. d. l., Hanna, E. B., Bressand, F., et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Kuperman, V. and Van Dyke, J. A. Effects of individual differences in verbal skills on eye-movement patterns during sentence reading. *Journal of Memory and Language*, 65(1):42–73, 2011.

Kuribayashi, T., Oseki, Y., Ito, T., Yoshida, R., Asahara, M., and Inui, K. Lower perplexity is not always human-like. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5203–5217, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.405. URL https://aclanthology.org/2021.acl-long.405.

Levy, R. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177, 2008. URL https://www.sciencedirect.com/science/article/abs/pii/S0010027707001436.

Linzen, T. and Jaeger, T. F. Uncertainty and expectation in sentence processing: Evidence from subcategorization distributions. *Cognitive science*, 40(6):1382–1411, 2016.

Nicenboim, B., Vasishth, S., Gattei, C., Sigman, M., and Kliegl, R. Working memory differences in long-distance dependency resolution. *Frontiers in Psychology*, 6:312, 2015.

Oh, B.-D. and Schuler, W. Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1915–1921, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.128. URL https://aclanthology.org/2023.findings-emnlp.128.

Oh, B.-D. and Schuler, W. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350, 2023b.

Pimentel, T., Meister, C., Wilcox, E. G., Levy, R. P., and Cotterell, R. On the effect of anticipation on reading times. *Transactions of the Association for Computational Linguistics*, 11:1624–1642, 2023.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9, 2019. URL https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

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

Schneider, G., Busse, B., Dumrukcic, N., and Kleiber, I. Do non-native speakers read differently? predicting reading times with surprisal and language models of native and non-native eye tracking data. *Language and Linguistics in a Complex World*, 32:153, 2023.

Sennrich, R., Haddow, B., and Birch, A. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL https://aclanthology.org/P16-1162.

Shain, C. CDRNN: Discovering complex dynamics in human language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3718–3734, 2021.

Shain, C., Meister, C., Pimentel, T., Cotterell, R., and Levy, R. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10):e2307876121, 2024.

Shannon, C. E. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. URL https://ieeexplore.ieee.org/document/6773024.

Škrjanec, I., Broy, F. Y., and Demberg, V. Expert-adapted language models improve the fit to reading times. *Procedia Computer Science*, 225:3488–3497, 2023.

Song, X., Salcianu, A., Song, Y., Dopson, D., and Zhou, D. Fast WordPiece tokenization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 2089–2103, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.160. URL https://aclanthology.org/2021.emnlp-main.160.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

van Schijndel, M. and Schuler, W. Approximations of predictive entropy correlate with reading times. In *39th Annual Meeting of the Cognitive Science Society*, pp. 1260–1265, 2017.

Vuong, L. C. and Martin, R. C. Domain-specific executive control and the revision of misinterpretations in sentence comprehension. *Language, Cognition and Neuroscience*, 29(3):312–325, 2014.

Wilcox, E., Meister, C., Cotterell, R., and Pimentel, T. Language model quality correlates with psychometric predictive power in multiple languages. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7503–7511, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.466. URL https://aclanthology.org/2023.emnlp-main.466.

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., and Levy, R. On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the 42th Annual Meeting of the Cognitive Science Society - Developing a Mind: Learning in Humans, Animals, and Machines, CogSci 2020*, 2020.

Wilcox, E. G., Pimentel, T., Meister, C., Cotterell, R., and Levy, R. P. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470, 2023b.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

A Details on predictors

A.1 Language Models

We deployed the following German LMs from the Huggingface library (Wolf et al., 2019):

- GPT-2 *base*: https://huggingface.co/benjamin/gerpt2
- GPT-2 *large*: https://huggingface.co/benjamin/gerpt2-large
- Llama 2 7b: https://huggingface.co/LeoLM/leo-hessianai-7b
- Llama 2 13b: https://huggingface.co/LeoLM/leo-hessianai-13b
- Mixtral: https://huggingface.co/mistralai/Mixtral-8x7B-v0.1

Lemma frequencies were extracted from dlexDB (Heister et al., 2011), based on the reference corpus underlying the Digital Dictionary of the German Language (DWDS; Berlin-Brandenburgische Akademie der Wissenschaften, 2016). *Word length* is defined as the number of characters including punctuation. Henceforth, we denote the word-level predictors surprisal s_i, contextual entropy h_i, log-lemma frequency f_i, and word length l_i for a word i.

A.2 Pooling of surprisal and contextual entropy to word level

We compute word-level surprisal by summing up the surprisal values of the individual sub-word tokens. Given k sub-word tokens $u_n, u_{n+1}, \ldots, u_{n+k}$ belonging to the same word token, the word token's surprisal is computed as

$$\begin{aligned} s(u_n, u_{n+1}, \ldots, u_{n+k}) &= -\log p(u_n, u_{n+1}, \ldots, u_{n+k} \mid \mathbf{u}_{<n}) \\ &= -\log \left[p(u_n \mid \mathbf{u}_{<n}) p(u_{n+1} \mid \mathbf{u}_{<n+1}) \ldots p(u_{n+k} \mid \mathbf{u}_{<n+k})\right] \\ &= -\log p(u_n \mid \mathbf{u}_n) + -\log p(u_{n+1} \mid \mathbf{u}_{<n+1}) + \cdots + -\log p(u_{n+k} \mid \mathbf{u}_{<n+k}), \end{aligned}$$

which shows that summing up sub-word token surprisal values is equivalent to computing the surprisal of the joint distribution of the sub-word tokens.

As regards entropy, we use the sum of the sub-word token-level contextual entropies as proxy for the joint entropy of the sub-word tokens' distribution. Given k $\bar{\Sigma}$-valued random variables $U_n, U_{n+1}, \ldots, U_{n+k}$ belonging to the same word token, their joint entropy is defined as:

$$\mathrm{H}(U_n, U_{n+1}, \ldots, U_{n+k}) \stackrel{\text{def}}{=} -\sum_{u_n \in \bar{\Sigma}} \sum_{u_{n+1} \in \bar{\Sigma}} \cdots \sum_{u_{n+k} \in \bar{\Sigma}} P(u_n, u_{n+1}, \ldots, u_{n+1}) \log_2 \left[P(u_n, u_{n+1}, \ldots, u_{n+1})\right].$$

However, depending on the tokenizer, the cardinality of $\bar{\Sigma}$ could be over 50,000, which makes the computation of the joint entropy computationally unfeasible. Instead, we use the sum of the individual entropies as proxy. This is only a proxy, since

$$\mathrm{H}(U_n, U_{n+1}, \ldots, U_{n+k}) \leq \mathrm{H}(U_n) + \mathrm{H}(U_{n+1}) + \cdots + \mathrm{H}(U_{n+k}).$$

This inequality is an equality iff $U_n, U_{n+1}, \ldots, U_{n+k}$ are statistically independent. Since this is not the case here, the sum of the sub-word token-level entropies is used as an upper bound.

B Individual Differences Corpus (InDiCo)

As mentioned in the main text, following previous work (Wilcox et al., 2023b, i.a.), we employ *first-pass reading time* (FPRT) –also referred to as *gaze duration*; the sum of all fixations on a word when fixating it for the first time–as a proxy for processing load: whereas total fixation duration can incorporate words from the right context due to regressive saccades, FPRT most strongly reflects the initial processing difficulty. Given that in our study, we only deploy auto-regressive LMs (cf.§4), FPRTs are also more in line with the fact that these models only have access to a word's left context.

We provide abbreviations and a brief summary of all psychometric tests in Tab. 2. More details can be found in Haller et al. (2023). A correlation matrix between all tests can be found in Fig.2. We can see strong correlations between many tests, in particular for the ones of the same psychological construct.

C Additional results

C.1 Baseline analyses

To corroborate results from previous work, we also assess the predictive power of entropy and surprisal in general, not taking into account individual psychometric scores. We define a baseline model $\mathcal{M}_0^b$ with predictors $\mathbf{x}_i^{b_0}$ including the

	Test / Measure	Construct	Description
Cognitive control	Stroop: reaction time effect	Verbal inhibitory cognitive control	Participants had to react (choose between congruent and incongruent) for color words whose font color either matched the content (congruent) or not (incongruent). Reaction time and accuracy were measured.
	Simon: reaction time effect	Non-verbal inhibitory cognitive control	Non-verbal equivalent to the Stroop Task where participants had to react (choose between congruent and incongruent) to arrows pointing to the right or left, shown either on the left or right side of the screen.
	FAIR: K score (total score)	Non-verbal cognitive control/attention	Participants had to find and mark target symbols (e.g.,, dice with 2 eyes among many other dice) on a page within a time limit. Measures of attentional performance, attention quality, and attention continuity were derived.
Working memory	Sentence span	Verbal working memory capacity	Participants had to judge the meaningfulness of sentences and remember letters presented after each sentence for later recall. In the end, they had to repeat all the letters.
	Operation span	Non-verbal working memory capacity	Participants were presented with consonants sequentially. After each consonant, they had to perform mathematical operations before the next consonant appeared. In the end, they had to repeat all consonants.
	Memory updating	Non-verbal working memory capacity	Participants had to remember an initial set of digits, each presented in a separate frame on the screen, and then update these digits in parallel through arithmetic operations.
	Spatial short-term memory	Non-Verbal Working Memory Capacity	Participants had to memorize the spatial locations of dots in a grid during a learning phase, and then locate them on an empty grid.
Intelligence	MWT: Percentile rank	Verbal intelligence/ word knowledge	Participants were presented lists of words, and for each list, they had to decide which of the presented words were real words.
	RIAS: verbal percentile rank	Verbal Intelligence	This test assessed verbal reasoning, and verbal logical thinking via question-answering and sentence completion.
	RIAS: non-verbal percentile rank	Non-Verbal Intelligence	This test assessed non-verbal reasoning and problem-solving tests where participants were presented with sets of images and they had to decide which image was not part of the set. In the other test, they had to identify missing elements in pictures.
	RIAS: total percentile rank	Intelligence	Total intelligence score based on verbal and non-verbal part.
Reading	SLRT: Word reading percentile rank	Reading fluency	Participants read out loud as many words within one minute as possible.
	SLRT: Pseudoword reading percentile rank	Reading fluency	Participants read out as many pseudo-words within one minute as possible.

Table 2: Psychometric tests conducted with all participants.

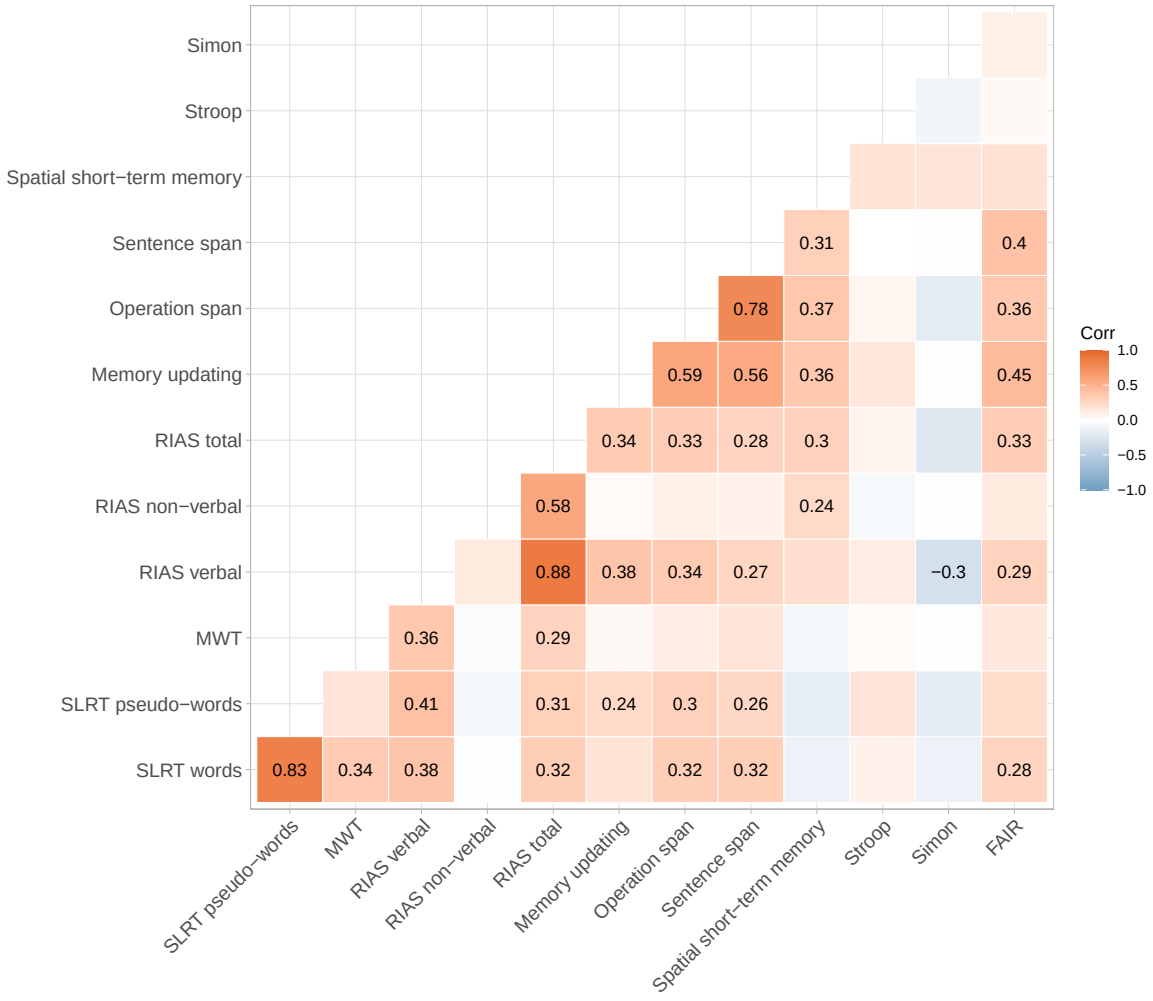

Figure 2: Correlations between scores of all psychometric tests. Red cells indicate positive correlation coefficients, blue cells negative correlation coefficients. Significant coefficients are displayed, blank cells indicate that the correlation was not significant with $\alpha = .05$.

word-level predictors word length l_i, log-lemma frequency f_i, a global intercept β_0, and an additional random by-subject intercept β_{0j}, *i.e.,*

$$\mathcal{M}_0^b : y_{ij} \sim \beta_0 + \beta_{0j} + \ \beta_1 \, l_i + \beta_2 \, f_i, \tag{3}$$

where y_{ij} refers to the log-transformed first-pass reading time of subject j for the i^{th} word in the stimulus corpus across all texts and following a log-normal distribution. The target models $\mathcal{M}_0^{t_s}$ and $\mathcal{M}_0^{t_h}$ solely include an additional surprisal or entropy term, *i.e.,* s_i or h_i.

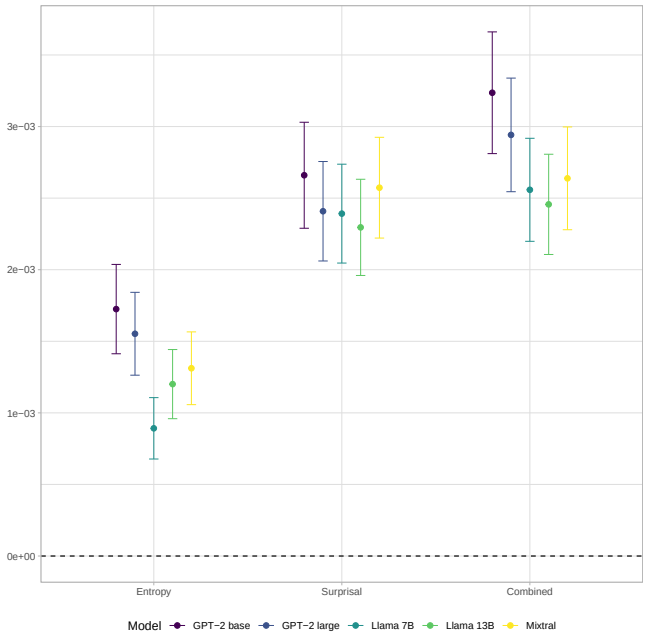

Figure 3: Predictive power of entropy and surprisal on reading times. Combined refers to the regression model where both predictors were included. Higher Δ_{LL} indicates higher predictive power.

As depicted in Fig.3 surprisal and contextual entropy exhibit predictive power (PP), albeit consistently lower for the latter. For GPT-2 *base* and *large*, adding both surprisal and contextual entropy as predictors increases the PP; for the other models, the combined version yields the same PP as using surprisal alone. Across models, GPT-2 *base* has the highest PP, with PP decreasing as model size increases.

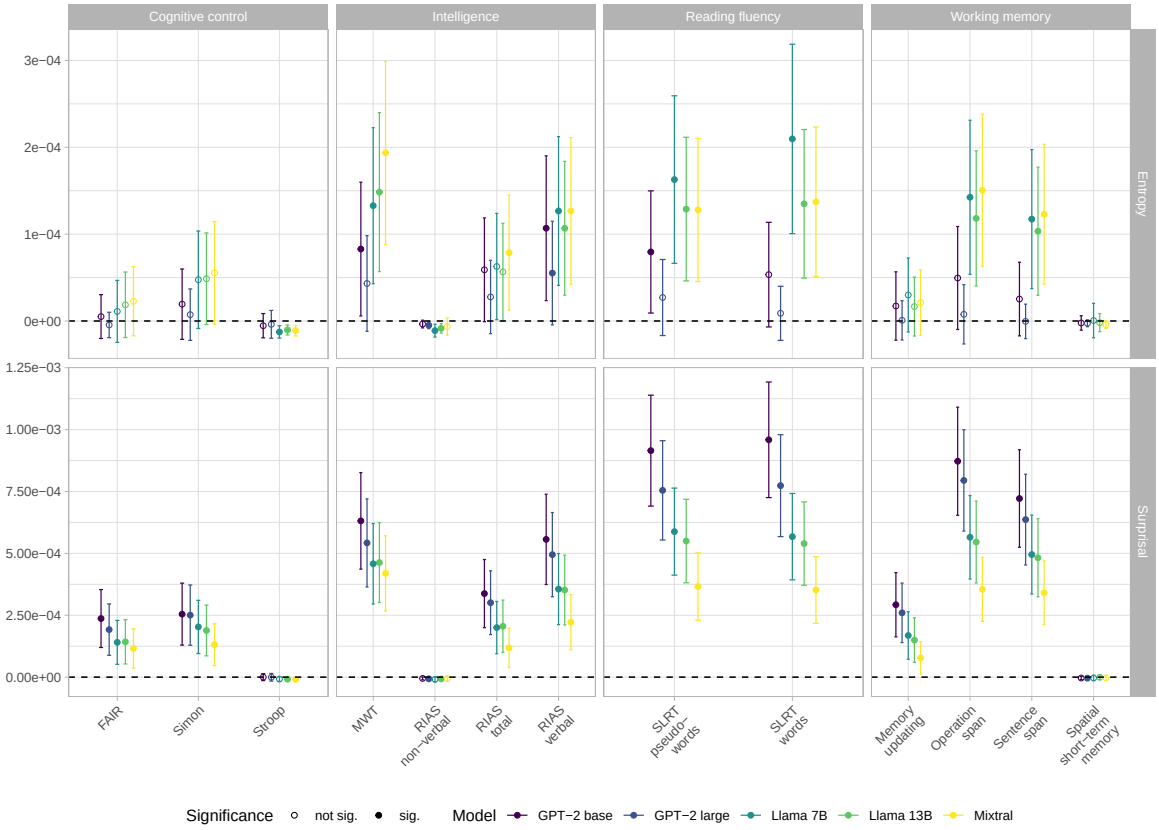

Figure 4: Δ_{LL} (mean and 95% CI) for the interactions between psychometric scores and model surprisal or entropy as additional predictors for reading times. Empty dots indicate that the Δ_{LL} is not significantly different from zero.