Language models emulate certain cognitive profiles: An investigation of how predictability measures interact with individual differences

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

[1]Department of Computational Linguistics, University of Zurich, Switzerland
[2]Department of Computer Science, University of Potsdam, Germany
{haller,bolliger,jaeger}@cl.uzh.ch

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 of surprisal and entropy measures estimated from a range of language models (LMs) on data of human reading times as a measure of processing effort by incorporating information of language users' cognitive capacities. To do so, we assess the predictive power of surprisal and entropy estimated from generative LMs on reading data obtained from individuals who also completed a wide range of psychometric tests. Specifically, we investigate if modulating surprisal and entropy relative to 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-performing groups, revealing what type of psycholinguistic subject a given LM emulates. Our study finds that in most cases, incorporating cognitive capacities increases predictive power of surprisal and entropy on reading times, and that generally, high performance in the psychometric tests is associated with lower sensitivity to predictability effects. Finally, our results 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.[1]

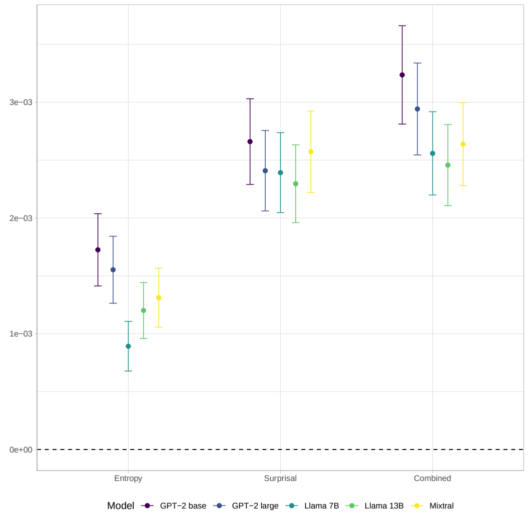

Figure 1: Predictive power of entropy and surprisal on reading times. Combined refers to the regression model where both predictors were included. Higher Δ_{LL} indicates higher predictive power.

1 Introduction

Human language comprehension and, by extension, human reading is incremental in nature: humans process words sequentially (Rayner and Clifton Jr, 2009), and different words in varying contexts impose different amounts of cognitive processing effort (Rayner, 1998). 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' probability distribution was operationalized by *surprisal theory* (Hale, 2001; Levy, 2008). Since then, a large body of research has investigated the exact nature of the relationship between surprisal and human processing effort, such as determining appropriate linking functions (Meister et al., 2021; Shain et al., 2024), or its manifestation in different languages (Wilcox et al., 2023a; Jäger et al., 2015; Kuribayashi et al., 2021, i.a.). Moreover, it has been shown repeatedly that both the quality of a model from which surprisal is extracted as well as the amount of data a model is trained on correlate with the model's psychometric predictive power[2], *i.e.,* its ability to predict human behavioral processing data (Frank and Bod, 2011; Fossum and Levy, 2012; Goodkind and Bicknell, 2018a), albeit

[1] Code is available at https://github.com/DiLi-Lab/LM-cog-profiles

[2] This hypothesis has been referred to as *quality-power hypothesis* by Wilcox et al. (2023a), cf. their introduction for a comprehensive summary.

Findings of the Association for Computational Linguistics: ACL 2024, pages 7878–7892
August 11-16, 2024 ©2024 Association for Computational Linguistics

only to a certain extent (Shain et al., 2024; Oh and Schuler, 2023b,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. However, an increasing body of research has demonstrated that an individual's characteristics do play an important role in human language processing (e.g., Estes, 1956; Daneman and Carpenter, 1980; Levinson, 2012; Van Dyke et al., 2014). Apart from surprisal, it has been suggested that reading processes are not purely *responsive*, which is mirrored in the surprisal effect of a word on a reader, but also *anticipatory*: readers make implicit assumptions about future words and allocate time to process it in advance. This anticipatory effect, which is reflected in the reading behavior, is induced by a reader's *expectation* about a word's surprisal and is operationalized as a word's *contextual entropy* (Linzen and Jaeger, 2016; van Schijndel and Schuler, 2017; Cevoli et al., 2022; Pimentel et al., 2023).

In this work, we revisit the relationship between predictability measures (surprisal and contextual entropy) and data of human processing effort in consideration of language users' individual cognitive differences. More specifically, we assess the *predictive power* (PP) of entropy, surprisal, as well as their interactions with cognitive measures, as predictors of human reading times in linear-mixed models. After establishing the baseline predictive power of surprisal and entropy on our German reading data ($\mathbf{H}_{\mathrm{B}}$), we investigate the following novel hypotheses:

$\mathbf{H}_1$: Modulating surprisal and entropy effects relative to individual cognitive capacities improves their predictive power on reading times on unseen data.

$\mathbf{H}_2$: Individuals with higher cognitive capacity rely less on predictive processing strategies, and hence exhibit lower surprisal or entropy effects.

$\mathbf{H}_3$: LMs are 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—GPT-2 base and large, Llama 2 7B and 13B, and Mixtral—to estimate both surprisal and contextual entropy and quantify their predictive power by including them as predictors in linear regressors, which are fitted to predict by-word reading times from InDiCo. If the regressors' log-likelihood improves after including these predictors and their interaction with the psychometric scores, we take this as corroboration of their predictive power.

We find that adding interaction terms between predictability measures (surprisal and entropy) and most cognitive scores significantly improves the quality of reading time predictions, and that in general, individuals with higher cognitive capacities exhibit smaller predictability effects. Lastly, there is evidence that LMs' abilities to predict reading times vary between high- and low-performing individuals within certain cognitive capacities. Specifically, all tested models emulate the processing behaviour of individuals with low verbal intelligence.

Our work is a first step towards investigating i) the differences in surprisal and entropy effects across different cognitive profiles, and ii) what type of cognitive biases might be inherent in the way LMs process language.

2 Related work

2.1 Surprisal and predictive power

Surprisal is a measure of predictability of a word in its context and has shown to be proportional to cognitive effort in human sentence processing (Hale, 2001; Levy, 2008). 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 its linear relationship with reading times (Demberg and Keller, 2008; Shain, 2021; Hoover et al., 2023; Pimentel et al., 2023), not just in English but also across languages (Pimentel et al., 2021; Wilcox et al., 2023a,b; de Varda and Marelli, 2022; Jäger et al., 2015; Kuribayashi et al., 2021). Moreover, researchers have investigated the degree to which surprisal is predictive of human reading times (i.e., assessing the *predictive power* (PP) of surprisal on human reading times) deploying different LMs. Wilcox et al. (2020) found that the better an LM's next-word expectation (*i.e.,* the lower its perplexity), the higher its PP. Along the same line, Wilcox et al. (2023a) demonstrated that an increasing LM quality, quantified by decreasing cross-entropy during training, leads to surprisal values that better predict RTs. Similarly, Goodkind and Bicknell (2018b) showed that the PP of surprisal increases linearly with the quality of LMs. This finding has since been refuted by Oh and Schuler (2023b) who revealed that large models, despite lower perplexity, provide worse PP of RTs. Oh and Schuler (2023a) further demonstrated that LMs provide the best fit to RTs when trained on around 2 billion tokens; beyond that point, additional training data causes the PP to decrease again.

2.2 Contextual entropy and predictive power

Linzen and Jaeger (2016) examined how sentence processing is affected by readers' uncertainty about the predictions they make during processing. They found that what they term single-step entropy (contextual entropy, cf.§3) does not affect RTs. However, they computed single-step entropy only over upcoming constituents based on verb subcategorization frames; it was later shown that entropy is indeed predictive of RTs (van Schijndel and Schuler, 2017). Cevoli et al. (2022) looked at the interaction of surprisal and entropy when predicting RTs: the impact of surprisal on reading behavior should vary as a function of entropy, such that surprising words inflict particularly high processing load when entropy is low. Wilcox et al. (2023b) examined whether contextual entropy is predictive of reading times and discovered that adding entropy as additional predictor (while keeping surprisal) increases PP, while replacing surprisal with entropy leads to a decrease in PP. However, Pimentel et al. (2023) also showed that using contextual entropy as a predictor in a linear-mixed model can be as good as surprisal when analyzing *anticipatory* effects reflected in word skipping rates, as opposed to *responsive* effects captured by gaze duration, for instance.

2.3 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, with evidence emerging that differences in cognitive abilities among language users do indeed have a significant impact on processing (Vuong and Martin, 2014; Nicenboim et al., 2015; Farmer et al., 2017, i.a.). For instance, Kuperman and Van Dyke (2011) demonstrated that measures related to cognitive control interact with word length and lexical frequency effects on fixation times, and Nicenboim et al. (2015) showed 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, in particular in the realm of native and non-native reading (Berzak and Levy, 2023; Schneider et al., 2023). For instance, Berzak and Levy (2023) demonstrated that higher L2 proficiency is associated with increased sensitivity to a word's predictability in context (surprisal). Moreover, Škrjanec et al. (2023) showed that specialized surprisal from domain-adapted LMs improves reading-time predictions for expert readers.

3 Methods

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

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

where $p(\cdot \mid \mathbf{u}_{<n})$ is the true distribution over words $u \in \bar{\Sigma}$ in context $\mathbf{u}_{<n}$. In other words, surprisal of a word is the negative log-probability conditioned on its left context.

Contextual entropy. The contextual entropy of a $\bar{\Sigma}$-valued random variable U_n at index n is the expected value of its surprisal, formalized as

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

It is a specific version of the Shannon entropy $\mathrm{H}(U) \overset{\text{def}}{=} -\sum_{u \in \mathcal{U}} p(u) \log p(u)$ that is conditioned on the left context of U (Shannon, 1948). As we do not have access to the true distribution $p(\cdot \mid \mathbf{u}_{<n})$, we approximate both measures using an auto-regressive language model p_θ.

3.1 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 want 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 $\mathcal{M}^b : \mathbf{x}_{ij}^b \mapsto y_{ij}$ that includes a set of baseline predictors $\mathbf{x}_{ij}^b$, and a target model $\mathcal{M}^t : \mathbf{x}_{ij}^b \oplus x_{ij}^q \mapsto y_{ij}$ that additionally includes the predictor of interest x_{ij}^q, where $\oplus$ represents the concatenation of two sets of predictors. Following previous work (Wilcox et al., 2020; Meister et al., 2021; Wilcox et al., 2023a; Pimentel et al., 2023, i.a.), we operationalize the predictive power as the mean difference in log-likelihood (Δ_{LL}) between the target and the baseline model, *i.e.*,

$$\Delta_{\mathrm{LL}} = \frac{1}{IJ}\Bigg[\sum_{i=1}^{I}\sum_{j=1}^{J}\log\mathcal{M}^t(y_{ij} \mid \mathbf{x}_{ij}^b \oplus x_{ij}^q) - \sum_{i=1}^{I}\sum_{j=1}^{J}\log\mathcal{M}^b(y_{ij} \mid \mathbf{x}_{ij}^b)\Bigg], \tag{3}$$

where I is the number of words and J is the number of subjects. 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 reading time data from InDiCo (Haller et al., 2023). This corpus contains eye-tracking-while-reading and self-paced-reading data from 61 native German speakers, collected across four experimental sessions, alongside a comprehensive battery of individual psychometric scores in four cognitive domains: cognitive control, working memory, intelligence, and reading fluency.[3] For our analyses, we use the standardized scores of 13 psychometric tests. 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.[4] 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.

Predictors

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. Crucially, we only consider auto-regressive LMs, as they most closely align with the incremental nature of human language comprehension (Hale, 2006; Rayner and Clifton Jr, 2009).

Since LMs employ tokenizers which split whitespace separated words into sub-word tokens (Sennrich et al., 2016; Song et al., 2021), word-level surprisal is computed by summing up the surprisal values of the sub-word tokens, which is equivalent to computing the surprisal of the joint distribution of sub-word tokens. 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.[5]

We further include lexical frequency and word length in our analyses since they are known to have an impact on human reading behavior. 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.

Psychometric scores. The psychometric assessment in InDiCo includes a total of 13 tests targeting different cognitive domains such as verbal and nonverbal working memory, cognitive control and intelligence, as well as reading fluency. A list of tests and their abbreviations can be found in Appendix B. For all test scores, higher scores originally indicate higher performance except for the *Stroop Reaction Time Effect* (Stroop) and the *Simon Reaction Time Effect* (Simon). In order to facilitate interpretability, we take the negative values of these scores such that a high value indicates high cognitive control. We standardize all scores in order to facilitate comparisons between tests. We denote the score of a

[3] For a detailed description of the tests, see Appendix B.

[4] Contrary to previous work, we do not set the reading time for words that were skipped to zero, but instead exclude them from the analysis. See limitations and caveats in Pimentel et al. (2023) on the influence of skipped words on surprisal estimation.

[5] For details on pooling of surprisal and entropy, see Appendix A.2.

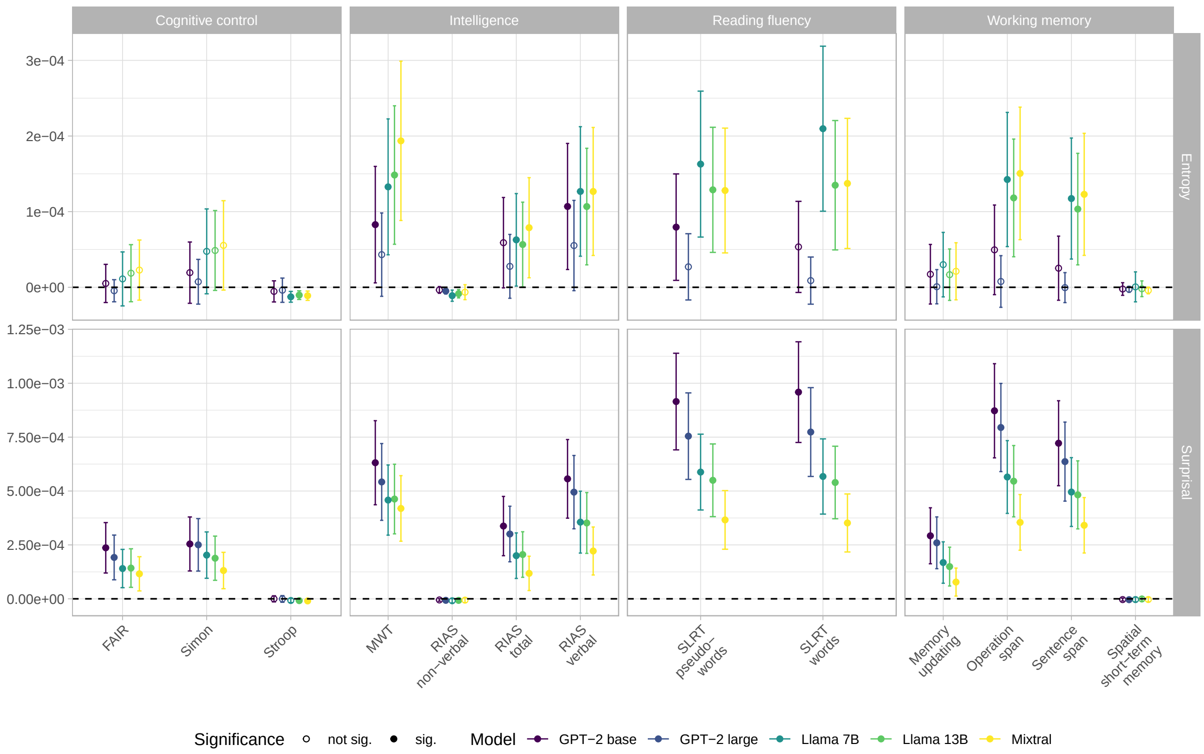

Figure 2: Δ_{LL} (mean and 95% CI) for the interactions between psychometric scores and model surprisal or entropy as additional predictors for reading times. Empty dots indicate that the Δ_{LL} is not significantly different from zero.

given psychometric test c for subject j.

4.1 Baseline analyses ($\mathbf{H_B}$)

To corroborate results from previous work, we first 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 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.,*

$$\mathcal{M}_0^b : y_{ij} \sim \beta_0 + \beta_{0j} + \ \beta_1 \, l_i + \beta_2 \, f_i, \tag{4}$$

where y_{ij} refers to the log-transformed first-pass reading time[6] 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.

Results. As depicted in Figure 1, surprisal and contextual entropy exhibit significant *predictive power* (PP), albeit consistently lower for entropy. 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.

4.2 Assessing the predictive power of interactions between surprisal/entropy and psychometric scores ($\mathbf{H_1}$)

To examine whether an interaction between cognitive scores and surprisal or entropy leads to an increase in predictive power on reading times, we define a baseline model $\mathcal{M}_1^b$ with predictors $\mathbf{x}_{ij}^{b_1}$ including the word-level predictors l_i, f_i, 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, and again a by-subject intercept β_{0j}, *i.e.,*

$$\begin{aligned}\mathcal{M}_1^b : y_{ij} \sim \beta_0 + \beta_{0j} + \ \beta_1 \, l_i + \beta_2 \, f_i + \\ \beta_3 \, s_i + \beta_4 \, h_i + \beta_5 \, c_j\end{aligned} \tag{5}$$

To assess whether allowing surprisal or entropy to be modulated by specific cognitive profiles—operationalized in terms of the individual psychometric measures—improves the prediction of reading time, we define target models $\mathcal{M}_1^{t_s}$ and $\mathcal{M}_1^{t_h}$ that include an additional interaction term between

[6] First-pass reading time denotes the sum of all fixation durations on a given word during its first pass (i.e., when reading it for the first time).

	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 memory	$0\ (\pm 0.001)^{\dagger}$	$0\ (\pm 0.001)$	$0.001\ (\pm 0.001)^{\dagger}$	$0.001\ (\pm 0.001)^{\dagger}$	$0.001\ (\pm 0.001)^{\dagger}$

Table 1: Effect sizes of interaction terms ± 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} (see Figure 2).

either surprisal or entropy and a given psychometric score c_j (e.g., *word-reading fluency score*) obtained for subject j, $x_{ij}^{q_1} \in \{s_i \cdot c_j, h_i \cdot c_j\}$:

$$\begin{aligned} \mathcal{M}_1^t : y_{ij} \sim \beta_0 + \beta_{0j} + \ & \beta_1\, l_i + \beta_2\, f_i + \\ & \beta_3\, s_i + \beta_4\, h_i + \beta_5\, c_j + \beta_6\, x_{ij}^{q_1} \end{aligned} \tag{6}$$

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 using the R library `broman` to establish whether a given Δ_{LL} is significantly different from 0 at $\alpha = .05$.

Results. Figure 2 shows the Δ_{LL} across all psychometric tests and models. Overall, we see that 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 extremely small) for these three scores across all models. Additionally, there are notable differences among different cognitive domains with respect to predictive power: modulating surprisal with scores targeting reading fluency or working-memory span yields the highest predictive power, followed by verbal intelligence scores. Scores targeting cognitive control show the lowest PP. The pattern observed in Figure 1 showing that increasing model size is associated with decreasing predictive power is visible here as well, but only for surprisal, not for entropy. Overall, interactions with surprisal extracted from the GPT-2 family have the highest PP. Conversely, interactions with GPT-2 based entropy have the lowest PP.

4.3 Assessing the magnitude of the interaction term coefficients (H$_2$)

To determine how specifically the surprisal and entropy effects are modulated by the psychometric scores, we run the target models $\mathcal{M}_1^{t_s}$ and $\mathcal{M}_1^{t_h}$ on the entire dataset and examine the effect sizes (coefficients) of the interaction term between the scores and the surprisal and entropy estimates, β_6. The coefficient of the interaction term indicates to what degree the fixed effect surprisal or entropy term is adjusted, relative to a subject's individual psychometric score, or, in other words, if individuals with a given cognitive profile are more sensitive to predictability effects. For instance, a positive coefficient for predictor $c_j \cdot s_j$ indicates that subjects with a higher score exhibit a stronger effect of surprisal, i.e. are more sensitive to a word's predictability, while a negative coefficient implies that subjects with a lower score exhibit a higher surprisal effect.

Results. We present the effect sizes of the interaction between scores and predictability measures in Table 1. First, we notice that for a given psy-

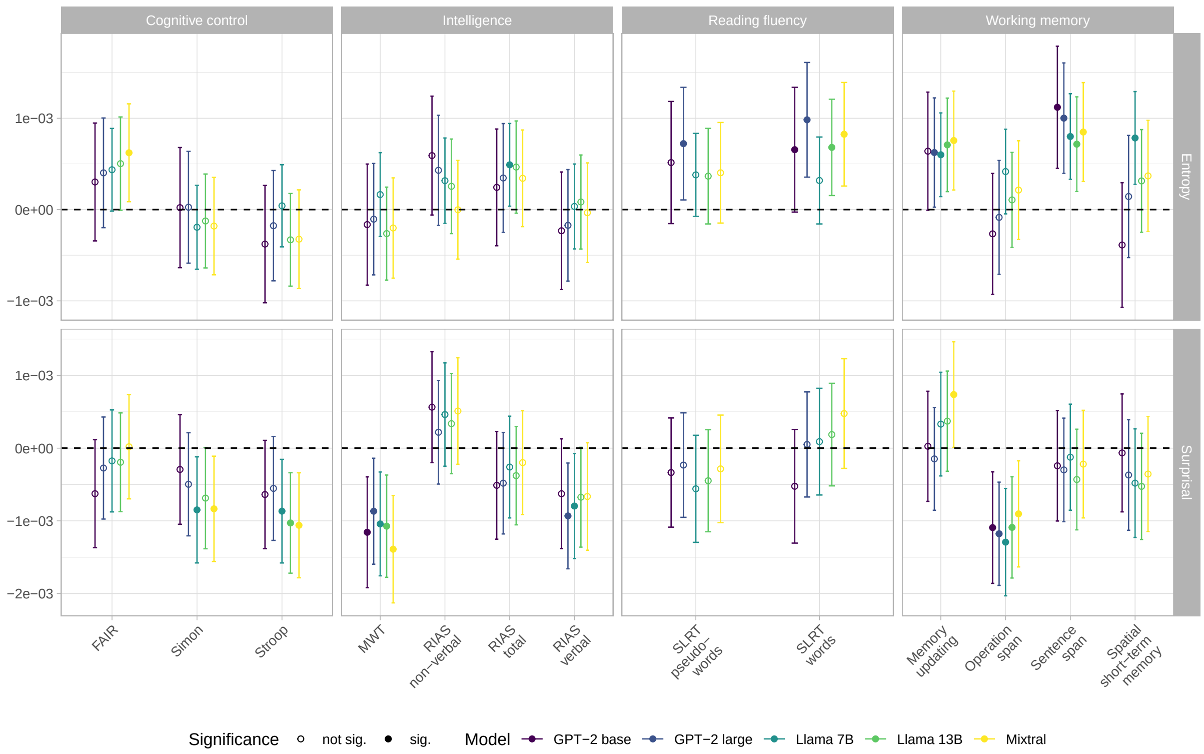

Figure 3: Difference in PP (ΔPP) (mean and 95% CI) of surprisal and contextual entropy for reading times. Positive ΔPP indicate higher PP for high-performing individuals; negative ΔPP higher PP for low-performing individuals. Empty dots indicate that the ΔPP is not significantly different from zero.

chometric 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. In these cases, individuals with higher scores show less sensitivity to a word's predictability (measured in terms of surprisal or entropy). 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.

4.4 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 (↑) and low-performing individuals (↓) at the median of each score. Then, for each group, we compute the Δ_{LL} between the baseline model $\mathcal{M}_3^b$ only including word length and lexical frequency as predictors 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 $\Delta_{\mathrm{LL}\downarrow}$ and $\Delta_{\mathrm{LL}\uparrow}$ 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 $\Delta\mathrm{PP} \overset{\text{def}}{=} \Delta_{\mathrm{LL}\uparrow} - \Delta_{\mathrm{LL}\downarrow}$. If a given LM is calibrated towards the cognitive profile of the high-performing individuals, we expect a positive ΔPP; and the ΔPP is negative if the LM is calibrated towards the low-performing cognitive profile.

Results. Figure 3 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. $\Delta\mathrm{PP} > 0$ indicates higher PP for the high-performing group, $\Delta\mathrm{PP} < 0$ 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. Regarding reading fluency, readers with high word-reading scores were significantly better predicted from entropy estimated via GPT-2 large, Llama 2 13B and Mixtral.

For surprisal, we find that across all models, RT predictions are significantly better for the the low-performing group in the operation span test as well as the vocabulary size test MWT. Moreover, surprisal extracted from GPT-2 large and Llama 2 7B leads to significant gains in PP for the low-performing group in the RIAS test, which like MWT assesses verbal intelligence. Finally, surprisal estimated from Llama 2 7B and Mixtral showed significantly higher PP for the group of individuals with lower scores in the verbal and non-verbal cognitive control tests (Stroop, Simon).

5 Discussion

Most studies on the predictive power of surprisal and entropy on reading times have been conducted on the group-level. Although individual-level effects may have been taken into account in terms of random slopes, these effects have not been associated with different cognitive profiles. Our findings suggest that (1) incorporating information on individuals' cognitive capacities and allowing them to modulate the magnitude of surprisal and entropy effects can increase the predictive power of these predictability measures, (2) individuals exhibit surprisal and entropy effects relative to certain cognitive capacities, and (3) some language models exhibit higher predictive power of reading times for groups of individuals associated with a certain cognitive profile.

5.1 Implications for the cognitive mechanisms of language processing

Fluent readers exhibit lower surprisal effects.[7] Our results show that the predictive power of the interaction terms between surprisal and reading fluency and to some extent entropy and reading fluency are particularly high compared to the interaction terms including other psychometric tests, as depicted in Figure 2. Including the interaction term between reading fluency scores and surprisal improves the predictions on reading times for all language models. The negative coefficient (Table 1) 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 4).

More accurate predictive processing for individuals with high working memory span. The results regarding the interaction terms between surprisal and working memory test scores are more difficult to contextualize. The span tests in particular (operation span, sentence span) lead to substantial increases in PP. Moreover, when assessing the magnitude of their interaction terms, we find that individuals with higher scores in both tests show lower surprisal effects as shown in Figure 2. At first glance, this intuitively makes sense, as high working memory can be associated with the capability to hold competing continuations in memory, including less likely ones that, in the high-surprisal situation, turn out to be the actual continuation. However, conversely, O'Rourke (2013) found in an ERP study that individuals with high operation span show stronger P600 effects, an event-related potential that is typically associated with syntactic repair or reanalysis. This would suggest that individuals with higher operation span exhibit a stronger garden path effect. However, garden paths are an extreme case of very high surprisal where different processing mechanisms might be at work such as re-analysis processes. Future experiments in minimal-pair settings will be needed to examine the connection between working memory capacity and surprisal effects more closely.

Attentiveness and inhibitory cognitive control may impact predictive processing differently. Next, we discuss the interaction term between surprisal and measures from tests targeting cognitive control. The directions of the interaction coefficients indicate that individuals with higher FAIR-scores (attention and concentration) exhibit weaker

[7] As many results for entropy are not significant or less clear, the focus of the discussion will lie on surprisal.

surprisal effects. Although this finding would be in line with the fact that general-purpose cognitive control mechanisms are required in the revision after linguistic misanalyses (Fedorenko, 2014), Vuong and Martin (2014) has shown that the time taken to revise a garden path is correlated only with verbal Stroop reaction time effects, but not with reaction time effects from its non-verbal counterpart (Simon). Since the FAIR scores showed fairly strong correlations with working-memory scores but not the other cognitive control tests (cf. Fig. 4), it is possible that the results obtained for FAIR might be more related to working-memory principles. Secondly, the results also showed that individuals performing well in the Simon test (inhibitory non-verbal cognitive control) exhibit stronger surprisal effects. As mentioned before, Vuong and Martin (2014) showed that the Simon task is likely not associated with mechanisms related to linguistic repair. The weaker surprisal effects for low-performing individuals in the Simon task is more likely associated with the tendency of participants with lower control to skip revising misinterpretations entirely, i.e., to rely on good-enough processing (Ferreira et al., 2002).

5.2 Cognitive profiles of language models

Finally, regarding the group analyses, the results presented in Figure 3 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 (un)certainty 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 3). For the high reading-fluency *pseudo-word reading* group, it even represents the only measure with a significant increase in PP. Taking into account that GPT-2 base and large showed a similar baseline PP (Figure 1), suggesting 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 aimed at uncovering model-internal biases, it might be worthwhile to, in turn, extend the investigation on 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, disregarding individual differences, assuming that the predictive power of next-word predictability metrics such as surprisal or entropy on human reading times is uniform across cognitive profiles. In this work, we have shown that indeed, LMs do exhibit systematic biases towards readers of certain cognitive profiles. This illustrates the usefulness of incorporating individual-level information within the study of LM interpretability and language modelling in general.

Limitations

Splitting the subjects at the median of their scores obtained in the respective cognitive tests is a straightforward way to split them into high- and low-performing groups and avoids the problem of class-imbalance. However, this kind of split might not group participants whose scores are distributed more narrowly. For future work, it might be sensible to utilize more sophisticated clustering approaches to obtain more cognitively homogeneous groups.

Moreover, recent work has shown that test-retest reliability of individual surprisal effects are low, *i.e.,* a surprisal effect for the same individual might vary on different days (Haller et al., 2023), depending on numerous factors such as wakefulness, motivation, but also random fluctuations. If this is true, we have to assume the predictive power with one and the same LM for a given subject, representing a cognitive profile, might be different depending on the subject's condition on that particular day. However, using the Indico data, this factor is controlled to some degree since it combines data from temporally separate sessions. That way, even if the surprisal effect does depend on external factors, merging the data from several sessions ensures more robust estimates of each subject's true surprisal effect. While it might still be a limitation, it is less so than for other conventional datasets. Finally, although InDiCo represents a fairly diverse sample in terms of age and gender, many participants have an academic background (see also Reich et al. (2024) on the necessity of including diverse populations in analyses of reading data).

Ethics Statement

Working with human data requires careful ethical considerations. The *Individual Differences Corpus* (InDiCo; Haller et al., 2023) utilized for this study follows the Helsinki Declaration (World Medical Association, 2013).

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.

References

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

Yevgeni Berzak and Roger Levy. 2023. Eye movement traces of linguistic knowledge in native and non-native reading. *Open Mind*, pages 1–18.

Benedetta Cevoli, Chris Watkins, and Kathleen Rastle. 2022. Prediction as a basis for skilled reading: insights from modern language models. *Royal Society Open Science*, 9(6):211837.

Meredyth Daneman and Patricia A. Carpenter. 1980. Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, 19(4):450–466.

Andrea de Varda and Marco Marelli. 2022. The effects of surprisal across languages: Results from native and non-native reading. In *Proceedings of the 2022 Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing (Findings)*, pages 138–144, Online only. Association for Computational Linguistics.

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

William K. Estes. 1956. The problem of inference from curves based on group data. *Psychological Bulletin*, 53(2):134–140.

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

Evelina Fedorenko. 2014. The role of domain-general cognitive control in language comprehension. *Frontiers in Psychology*, 5:39895.

Fernanda Ferreira, Karl GD Bailey, and Vittoria Ferraro. 2002. Good-enough representations in language comprehension. *Current Directions in Psychological Science*, 11(1):11–15.

Victoria Fossum and Roger Levy. 2012. Sequential vs. hierarchical syntactic models of human incremental sentence processing. In *Proceedings of the 3rd workshop on cognitive modeling and computational linguistics (CMCL 2012)*, pages 61–69.

Stefan L Frank and Rens Bod. 2011. Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, 22(6):829–834.

Adam Goodkind and Klinton Bicknell. 2018a. 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.

Adam Goodkind and Klinton Bicknell. 2018b. 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.

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*.

John Hale. 2006. Uncertainty about the rest of the sentence. *Cognitive Science*, 30(4):643–672.

Patrick Haller, Iva Koncic, David Reich, and Lena A Jäger. 2023. Measurement reliability of individual differences in sentence processing: A cross-methodological reading corpus and bayesian analysis. *ArXiv Preprint*.

Julian Heister, Kay-Michael Würzner, Johannes Bubenzer, Edmund Pohl, Thomas Hanneforth, Alexander Geyken, and Reinhold Kliegl. 2011. dlexdb–eine lexikalische datenbank für die psychologische und linguistische forschung. *Psychologische Rundschau*.

Jacob Louis Hoover, Morgan Sonderegger, Steven T Piantadosi, and Timothy J O'Donnell. 2023. The plausibility of sampling as an algorithmic theory of sentence processing. *Open Mind*, 7:350–391.

Lena Jäger, Zhong Chen, Qiang Li, Chien-Jer Charles Lin, and Shravan Vasishth. 2015. The subject-relative advantage in chinese: Evidence for expectation-based processing. *Journal of Memory and Language*, 79:97–120.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

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

Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, and Kentaro Inui. 2021. Lower perplexity is not always human-like. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5203–5217, Online. Association for Computational Linguistics.

Stephen C. Levinson. 2012. The original sin of cognitive science. *Topics in Cognitive Science*, 4(3):396–403.

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

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

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the Uniform Information Density hypothesis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 963–980, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bruno Nicenboim, Shravan Vasishth, Carolina Gattei, Mariano Sigman, and Reinhold Kliegl. 2015. Working memory differences in long-distance dependency resolution. *Frontiers in Psychology*, 6:312.

Byung-Doh Oh and William Schuler. 2023a. Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Findings)*, pages 1915–1921, Singapore. Association for Computational Linguistics.

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

Polly O'Rourke. 2013. The interaction of different working memory mechanisms and sentence processing: A study of the p600. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 35.

Tiago Pimentel, Clara Meister, Elizabeth Salesky, Simone Teufel, Damián Blasi, and Ryan Cotterell. 2021. A surprisal–duration trade-off across and within the world's languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Tiago Pimentel, Clara Meister, Ethan G Wilcox, Roger P Levy, and Ryan Cotterell. 2023. On the effect of anticipation on reading times. *Transactions of the Association for Computational Linguistics*, 11:1624–1642.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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

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

David R. Reich, Shuwen Deng, Marina Björnsdóttir, Lena Jäger, and Nora Hollenstein. 2024. Reading does not equal reading: Comparing, simulating and exploiting reading behavior across populations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13586–13594, Torino, Italia. ELRA and ICCL.

Gerold Schneider, Beatrix Busse, Nina Dumrukcic, and Ingo Kleiber. 2023. 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.

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. 2021. 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)*, pages 3718–3734.

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.

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

Iza Škrjanec, Frederik Yannick Broy, and Vera Demberg. 2023. Expert-adapted language models improve the fit to reading times. *Procedia Computer Science*, 225:3488–3497.

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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Julie A. Van Dyke, Clinton L. Johns, and Anuenue Kukona. 2014. Low working memory capacity is only spuriously related to poor reading comprehension. *Cognition*, 131(3):373–403.

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

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

Ethan Wilcox, Clara Meister, Ryan Cotterell, and Tiago Pimentel. 2023a. Language model quality correlates with psychometric predictive power in multiple languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7503–7511, Singapore. Association for Computational Linguistics.

Ethan G Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P Levy. 2023b. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

World Medical Association. 2013. World medical association declaration of helsinki: ethical principles for medical research involving human subjects. *Jama*, 310(20):2191–2194.

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

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:

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

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)

We provide abbreviations and a brief summary of all psychometric tests in Table 2. More details can be found in Haller et al. (2023). A correlation matrix between all tests can be found in Figure 4. We can see strong correlations between many tests, in particular for the ones of the same psychological construct.

	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.

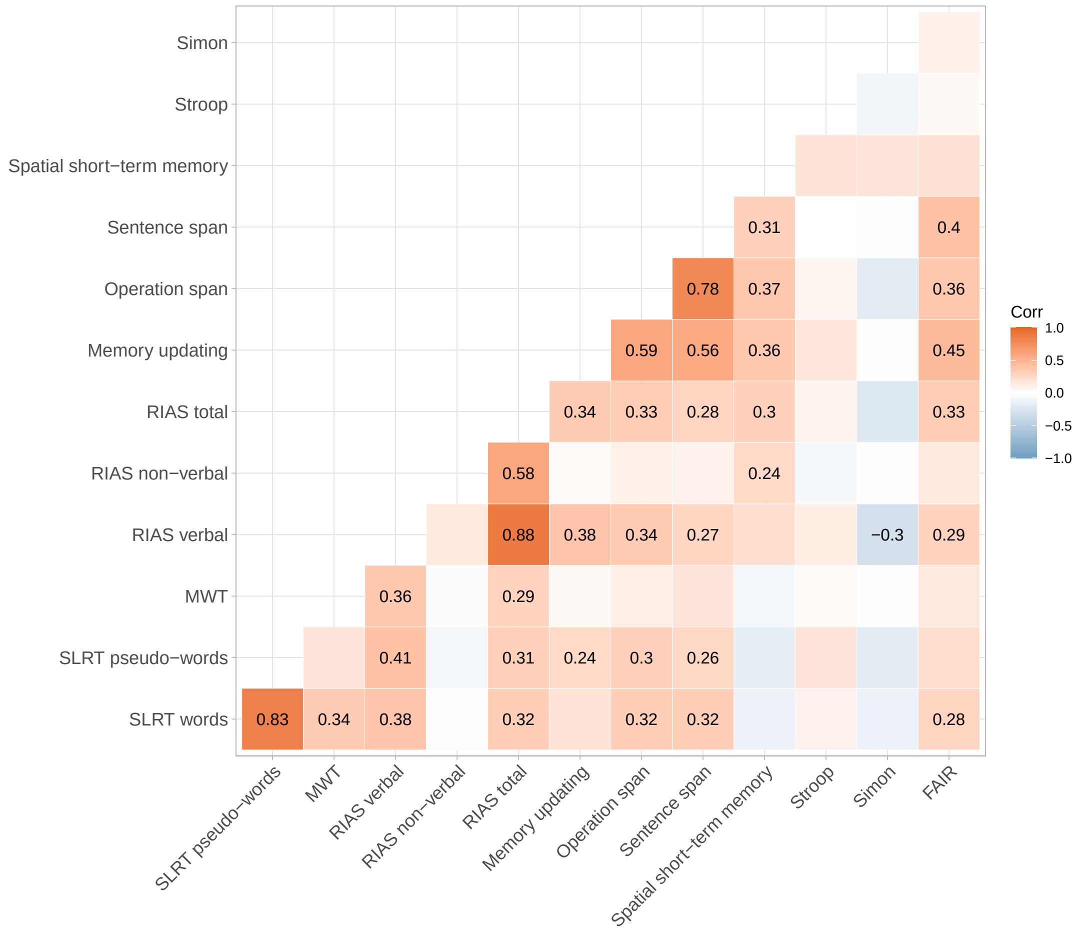

Figure 4: Correlations between scores of all psychometric tests. Red cells indicate positive correlation coefficients, blue cells negative correlation coefficients. Significant coefficients are displayed, blank cells indicate that the correlation was not significant with $\alpha = .05$.