Language models' probability distributions are calibrated to cognitive profiles: an investigation of the predictive power of surprisal and entropy

Anonymous ACL submission

Abstract

To date, most investigations on predictability (surprisal and entropy) effects have been conducted on the group-level, disregarding individual differences. In this work, we revisit the predictive power of different LMs' predictability effects on data of human processing effort by considering information of language users' cognitive capacities. To do so, we assess the predictive power of predictability effects estimated from generative LMs on reading data from subjects for which we also scores from psychometric tests targeting different cognitive domains. We investigate if allowing the model to modulate predictability effects relative to the cognitive scores increases the prediction of reading times, and we examine whether LMs exhibit systematic biases in the prediction of reading times for cognitively high- or low-performing groups, allowing us to reveal what type of psycholinguistic subjects a given LM emulates. Our results confirm that incorporating cognitive capacities and allowing them to modulate surprisal and entropy effects can increase the predictive power of these predictors. Moreover, we find that predictive power of surprisal extracted from GPT-2 degrades for individuals with higher verbal intelligence but also improves for individuals with higher verbal cognitive control, suggesting that GPT-2 emulates a subject with high verbal cognitive control but low verbal intelligence.

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 changing contexts impose different amounts of cognitive processing efforts. The relationship between a language that is probabilistic in nature (Grenander, 1967; Wetherell, 1980) and cognitive effort was operationalized as *surprisal* (Hale, 2001; Levy, 2008). Ever since the introduction of surprisal theory, a large body of research has investigated the details of the relationship between surprisal and human processing effort, such as determining appropriate linking functions (Meister et al., 2021; Shain et al., 2022), or its manifestation in different languages (Pimentel et al., 2021; Wilcox et al., 2023a,b; de Varda and Marelli, 2022). 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[1], *i.e.,* its ability to predict human behavioral processing data (Frank and Bod, 2011; Fossum and Levy, 2012; Goodkind and Bicknell, 2018a), albeit only to a certain extent (Shain et al., 2022; Oh and Schuler, 2023b,a). So far, most studies have tested these predictions on the group-level, assuming that a person's individual properties are independent of the predictions un-

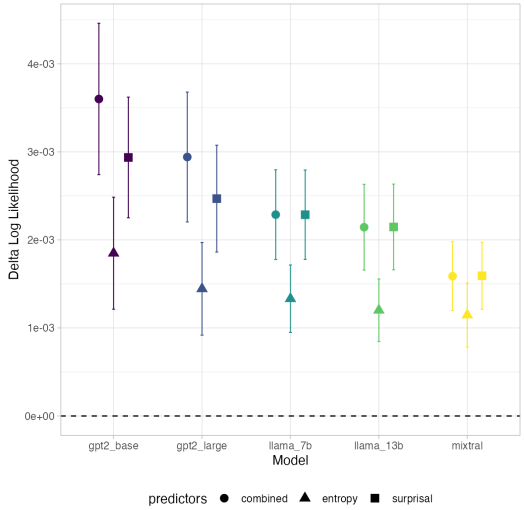

Figure 1: Predictive power of surprisal and entropy, extracted from different language models, on reading times, used as separate and combined (*i.e.,* both included in the same linear model) predictors. Higher Δ_{LL} indicates higher predictive power.

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

der surprisal theory. 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 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, whether the predicted individual effect is similar across different language models (LMs), and we analyse the impact of individual cognitive differences on the psychometric predictive power of surprisal and entropy estimated from a range of LMs. To that end, we investigate the following hypotheses:

- $\mathbf{H}_0$: The inclusion of surprisal/entropy leads to a higher predictive power on reading times.
- $\mathbf{H}_1$: Adjusting surprisal/entropy effects relative to individual cognitive capacities improves the predictive power on reading times on unseen data.
- $\mathbf{H}_2$: Individuals with scores indicating higher cognitive performance experience a lower surprisal/entropy effect as they are expected to make more precise predictions about upcoming material.
- $\mathbf{H}_3$: The predictive power of surprisal/entropy differs between groups of different 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 psychometric assessment targeting various cognitive capacities. We deploy a range of pre-trained generative LMs to estimate both surprisal and contextual entropy and quantify their predictive power by including them as predictors in linear regressors, which are trained 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 effects (surprisal and entropy) and cognitive scores significantly impacts the quality of reading time predictions, and that in general, individuals with higher cognitive capacities experience lower predictability effects. Lastly, there is evidence that LMs's abilities to predict reading times vary between individuals with specific cognitive abilities. Specifically, most models emulate the processing behaviour of high-working memory and low-verbal intelligence. Importantly, all these findings are consistent across a range of LMs.

Our work is a first step towards investigating whether it might be worthwhile to consider information on individual cognitive capacities for language modeling in general.

2 Related work

2.1 Surprisal and predictive power

Surprisal (Hale, 2001; Levy, 2008) is a measure of predictability of a word and said 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 psychometric measurements in natural reading (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). More recently, research has focused on the predictive power (PP) of LMs that estimate surprisal on human processing data. For instance, Wilcox et al. (2020a) find that the better an LM's next-word expectation (*i.e.,* the lower its perplexity), the better is its PP. Along the same line, Wilcox et al. (2023a) demonstrate 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) show that the PP of surprisal increases linearly with the quality of LMs. This finding has since been refuted by Oh and Schuler (2023b) who reveal that large models, despite their lower perplexity, underpredict RTs. Oh and Schuler

(2023a) further demonstrate 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) examine how sentence processing is affected by readers' uncertainty about the predictions they make during processing. They find that what they term single-step entropy (contextual entropy) does not affect RTs. However, they compute the single-step entropy only over upcoming constituents based on verb subcategorization frames, and later van Schijndel and Schuler (2017) show that entropy is indeed predictive: there is a positive correlation between predictive entropy and RTs. Cevoli et al. (2022) look 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 are particularly detrimental when entropy is low. Wilcox et al. (2023b) examine whether contextual entropy is predictive of reading times and discover that adding entropy as additional predictor (while also keeping surprisal) improves the model's PP, while replacing surprisal with entropy hurts the PP. This result is, however, not substantiated by Pimentel et al. (2023), who show that the reading process is both responsive and anticipatory: in most cases, adding only contextual entropy as predictor is as good as adding surprisal.

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) 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, 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) 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

Surprisal. 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 $\mathbf{u}_{<n}$. In other words, surprisal is the negative log-probability conditioned on its left context. As we do not have access to the true distribution $p(\cdot \mid \mathbf{u}_{<n})$, we approximate it using an auto-regressive language model p_θ.

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

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

It is a specific version of the Shannon entropy (Shannon, 1948) $\mathrm{H}(U) \stackrel{\text{def}}{=} -\sum_{u \in \mathcal{U}} p(u) \log p(u)$ that is conditioned on the left context of U. Again, since we do not know the true distribution $p(\cdot \mid \mathbf{u}_{<n})$, we approximate the conditional entropy using a 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 for a subject j on word i, from a set of 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 $\mathcal{M}^b : \mathbf{x}^b_{ij} \mapsto y_{ij}$ 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., 2020b; 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.*,

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

where I is the number of sentences 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[2]

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 41 native German speakers, collected across four sessions, alongside 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*–also called *gaze duration*; the sum of all fixations on a word when passing it the first time–as a proxy for processing load: whereas total fixation duration can incorporate words from the right context due to regressive saccades, *first-pass reading time* most strongly reflects the first-pass processing difficulty. 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.

4.1 Predictors

Word-level predictors To extract surprisal and contextual entropy, we deploy the German versions of a range of pretrained transformer-based LMs, 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 autoregressive LMs, as they most closely align with human language comprehension and in turn reading behavior, which is assumed to be incremental (Hale, 2006; Rayner and Clifton Jr, 2009).

Since LMs employ tokenizers which split word tokens 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.[4]

Other predictors we include in our analyses and that are known to have an impact on human reading behavior are lexical frequency and word length. Lexical frequencies were manually extracted from the lexical database dlexDB (Heister et al., 2011), based on the reference corpus underlying the Digital Dictionary of the German Language (DWDS) (Das Digitale Wörterbuch der deutschen Sprache), and manually corrected. *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 non-verbal working memory, cognitive control, reading fluency, and intelligence. A list of tests and their abbreviations can befound in Appendix B. Note that for all test scores, higher scores indicate higher performance except for the *Stroop Reaction Time Effect* (StrRTEffect) and the *Simon Reaction Time Effect* (SimRTEffecT). In order to facilitate interpretability, we flip their scores such that a high value indicates high cognitive control. We standardize (z-transform) all scores in order to facilitate comparisons between tests. We denote the score of a given psychometric test c_j for subject j.

4.2 Baseline analyses ($\mathbf{H_0}$)

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 and an additional 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 first-pass reading time of subject j for the ith word in the stimulus corpus

[2] Our code is available at `anonymous-link`.

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

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

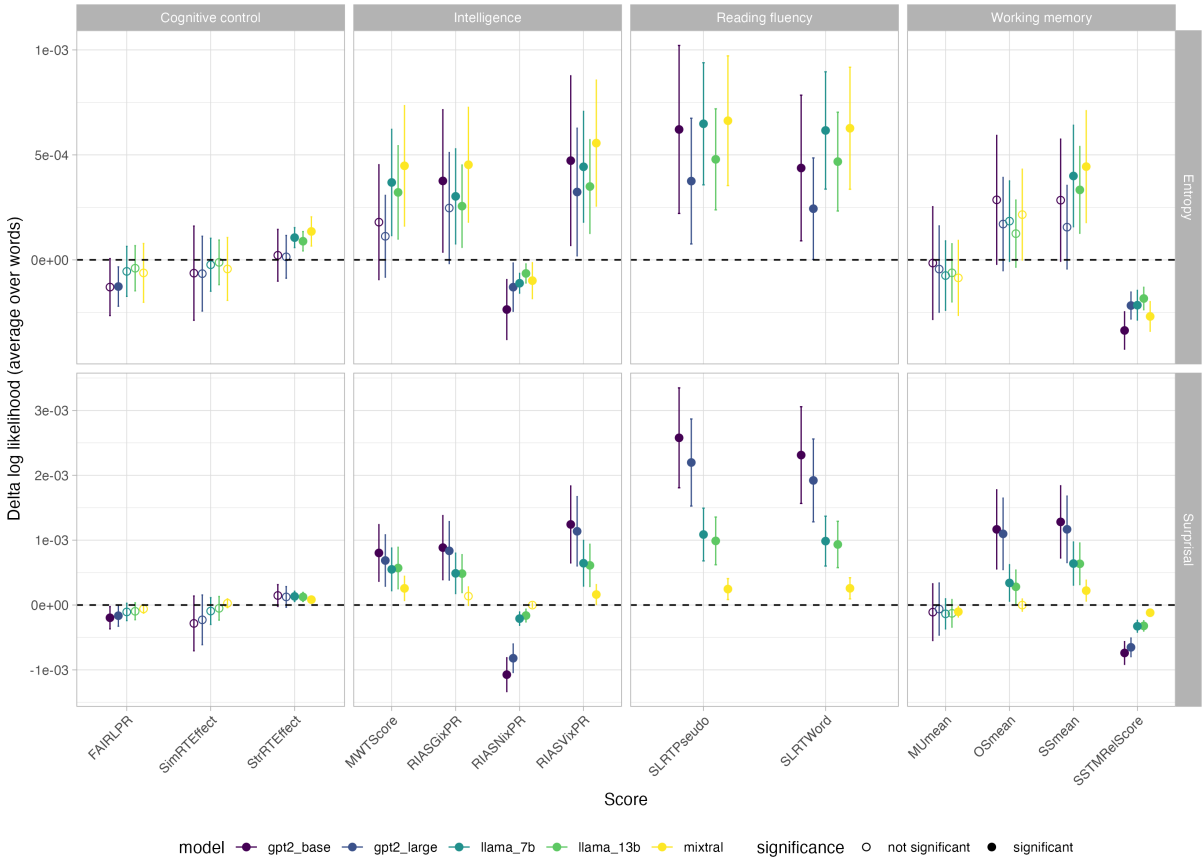

Figure 2: Average Δ_{LL} and 95% confidence intervals with the interactions between psychometric scores and model surprisal or entropy, extracted from different LMs, as additional predictors for reading times. Empty dots indicate that the Δ_{LL} is not significantly different from zero.

across all texts. β_0 represents the global intercept, and β_{0j} the random intercept for subject j, with y_{ij} following a 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 both surprisal and contextual entropy have predictive power (PP), although it is consistently lower for the latter. For both 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.

4.3 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 of 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, f_i, s_i, h_i, and the subject-level predictor c_j denoting a 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.,*

$$\mathcal{M}_1^b : y_{ij} \sim \beta_0 + \beta_{0j} + \beta_1\, l_i + \beta_2\, f_i + \beta_3\, s_i + \beta_4\, h_i + \beta_5\, c_j. \quad (5)$$

To assess whether allowing surprisal or entropy to be modulated by specific cognitive profiles—determined by 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 either surprisal or entropy and a given psychometric score c_j (e.g., *word-reading fluency score*) obtained for subject j, $x_{ij}^{q_1} \in \{s_i * c_j, h_i * c_j\}$:

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

A positive Δ_{LL} between the target and the baseline model indicates that a given psychometric test improves the prediction on the held-out test data.

Cognitive domain	Test	Effect size of interaction term				
		GPT-2 *base*	GPT-2 *large*	Llama-2 7b	Llama-2 13b	Mixtral
Cognitive control	FAIRLPR	-0.6 (±0.1)	-0.64 (±0.11)	-0.49 (±0.08)†	-0.49 (±0.08)†	-0.18 (±0.05)†
	StrRTEffect	-0.78 (±0.13)†	-0.81 (±0.13)†	-0.36 (±0.11)	-0.31 (±0.11)	-0.04 (±0.07)
	MWTScore	-1.55 (±0.1)	-1.52 (±0.1)	-1.13 (±0.08)	-1.12 (±0.08)	-0.48 (±0.05)
Intelligence	RIASVixPR	-2.12 (±0.1)	-2.11 (±0.11)	-1.26 (±0.08)	-1.2 (±0.08)	-0.41 (±0.05)†
	RIASNixPR	-0.24 (±0.1)	-0.28 (±0.11)	-0.2 (±0.09)	-0.23 (±0.08)	-0.09 (±0.05)†
	RIASGixPR	-1.69 (±0.09)	-1.7 (±0.1)	-1.04 (±0.08)	-1.01 (±0.08)	-0.36 (±0.05)
Reading fluency	SLRTWord	-2.56 (±0.1)	-2.43 (±0.1)	-1.38 (±0.08)	-1.32 (±0.08)	-0.45 (±0.05)
	SLRTPseudo	-2.63 (±0.1)	-2.54 (±0.1)	-1.45 (±0.08)	-1.36 (±0.08)	-0.44 (±0.05)
	MUmean	-1.36 (±0.1)†	-1.39 (±0.1)†	-0.73 (±0.08)†	-0.68 (±0.08)†	-0.17 (±0.05)
Working memory	OSmean	-2.32 (±0.11)	-2.32 (±0.11)	-1.11 (±0.09)	-1.02 (±0.09)	-0.27 (±0.06)†
	SSmean	-2.02 (±0.1)	-2.01 (±0.1)	-1.2 (±0.08)	-1.17 (±0.08)	-0.43 (±0.05)
	SSTMRelScore	0.57 (±0.11)	0.46 (±0.11)	0.28 (±0.09)	0.25 (±0.09)	0.1 (±0.05)

Table 1: Effect sizes of interaction terms ± standard error between surprisal, extracted from different LMs and psychometric test scores. The dagger † indicates that the inclusion of the interaction terms did not lead to a significant increase or decrease improve model performance (see Figure 2).

Cognitive domain	Test	Effect size of interaction term				
		GPT-2 *base*	GPT-2 *large*	Llama-2 7b	Llama-2 13b	Mixtral
Cognitive control	FAIRLPR	-0.75 (±0.17)†	-0.55 (±0.18)	-0.6 (±0.11)†	-0.58 (±0.11)†	-0.74 (±0.12)†
	StrRTEffect	-0.84 (±0.21)†	-0.81 (±0.22)†	-0.28 (±0.14)	-0.29 (±0.14)	-0.46 (±0.15)
	MWTScore	-1.36 (±0.17)†	-1.14 (±0.18)†	-1.21 (±0.11)	-1.13 (±0.11)	-1.44 (±0.12)
Intelligence	RIASVixPR	-2.26 (±0.17)	-1.98 (±0.18)	-1.34 (±0.11)	-1.22 (±0.11)	-1.64 (±0.12)
	RIASNixPR	-0.51 (±0.17)	-0.5 (±0.18)	-0.1 (±0.11)	-0.17 (±0.11)	-0.35 (±0.12)
	RIASGixPR	-1.89 (±0.16)†	-1.68 (±0.17)	-1.07 (±0.1)	-1.01 (±0.11)	-1.39 (±0.11)
Reading fluency	SLRTWord	-1.95 (±0.17)	-1.5 (±0.18)	-1.41 (±0.11)	-1.26 (±0.11)	-1.57 (±0.12)
	SLRTPseudo	-2.24 (±0.16)	-1.85 (±0.17)	-1.45 (±0.11)	-1.28 (±0.11)	-1.63 (±0.12)
Working memory	SSmean	-1.68 (±0.17)†	-1.33 (±0.18)†	-1.21 (±0.11)	-1.1 (±0.11)	-1.4 (±0.12)
	SSTMRelScore	0.31 (±0.18)	0.16 (±0.19)	0.32 (±0.12)	0.23 (±0.12)	0.27 (±0.13)

Table 2: Effect sizes of interaction terms ± standard error between entropy, extracted from different LMs and psychometric test scores. The dagger † indicates that the inclusion of the interaction terms did not lead to a significant increase or decrease model performance (see Figure 2).

Results. Figure 2 shows the Δ_{LL} across all psychometric tests and models. We see that in general, adding surprisal results in a higher increase in predictive power compared to entropy. However, by and large, the model- and score-specific Δ_{LL} follow similar patterns (i.e., consistency in terms of direction and magnitude of the Δ_{LL}): for instance, for both surprisal and entropy, the non-verbal RIAS score has the lowest Δ_{LL} out of all intelligence test scores. Moreover, there are notable differences among different cognitive domains with respect to predictive power: modulating surprisal with scores targeting reading fluency yields the highest predictive power, followed by scores for intelligence (with the exception of RIASNixPR), while scores for cognitive control do not help in improving reading time prediction. The pattern revealed in the Figure 1 for H_0 that increasing model size is associated with decreasing predictive power can also be seen, but only for surprisal. Overall, though, metrics extracted from the GPT-2 family seem to have the strongest interactions with most scores, despite a few exceptions.

4.4 Assessing effect sizes (H_2)

To determine how specifically the surprisal and entropy effects are modulated by the cognitive 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 direction of the effect indicates whether a given model has to adjust its own surprisal estimates relative to a subject's individual psychometric score in order to optimize its reading time predictions, or, in other words, whether a model's global surprisal over- or underestimates individual's surprisal effects. For instance, a positive effect for predictor c_j indicates that subjects with a higher score experience a higher effect of surprisal or contextual entropy, while a negative effect implies that subjects with a lower score experience a higher surprisal or contextual entropy effect.

Results. We present the effect sizes of the interaction between scores and predictability measures in Table 1 and Table 2. We see that for most psychometric tests, higher scores result in a reduction

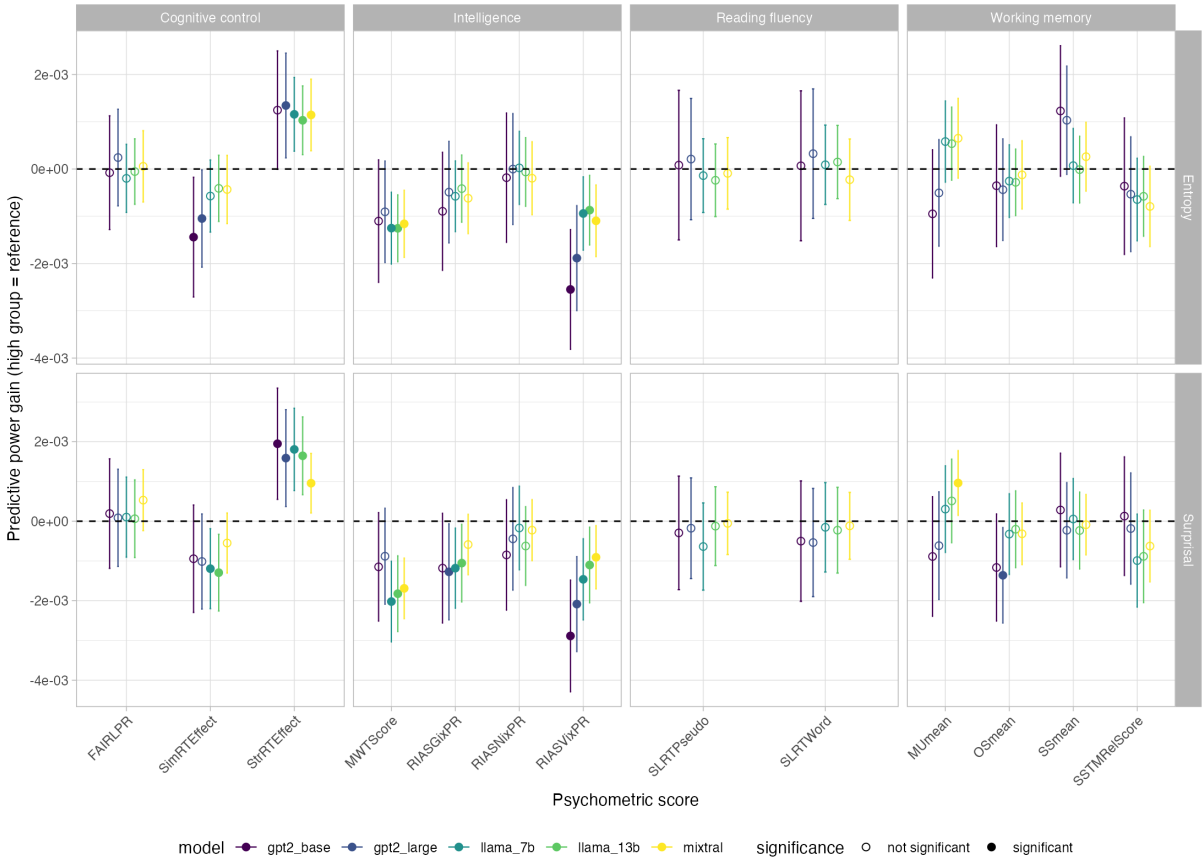

Figure 3: Predictive power gain (PPG) and 95% confidence intervals for surprisal and and contextual entropy predicting RTs for the high-performing subgroup, extracted from different LMs, for a range of psychometric tests. A positive PPG indicates higher PP for high-performing individuals; a negative PPG for low-performing individuals. Empty dots indicate that the Δ_{LL} is not significantly different from zero.

of the surprisal effect, indicated by the fact that the interaction term effects are negative: individuals with higher scores experience lower effects for both surprisal and entropy for a given word with fixed surprisal or entropy, respectively. This holds true across all models. Notably, for a given psychometric test, all models consistently modulate surprisal and entropy effects in the same direction.

4.5 Assessing the gain or loss of predictive power by sub-groups ($\mathbf{H_3}$)

Lastly, we investigate whether there are differences in how well language models' surprisal and entropy estimates are calibrated towards certain cognitive profiles, or, in other words, what type of psycholinguistic subject a given language model emulates. To do so, we split the reading time data at the median of the scores of each respective test to create a subset of data from a high-performing (↑) versus low-performing individuals (↓), relative to 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 $x_i^{q_3} \in \{s_i, h_i\}$. The individual $\Delta_{\text{LL}\downarrow}$ and $\Delta_{\text{LL}\uparrow}$ indicate the predictive power of the predictor of interest for each group separately. In order to answer which group experienced a higher relative gain in predictive power, we define the predictive power gain $\text{PPG} \stackrel{\text{def}}{=} \Delta_{\text{LL}\uparrow} - \Delta_{\text{LL}\downarrow}$. If a given LM is calibrated well towards the cognitive profile of the high-performing individuals, we expect a a positive PPG; and the PPG is negative if it is calibrated well towards the low-performing cognitive profile. This measure accounts for the fact that predictions might be better for a given group *per se* and indicates whether surprisal or entropy extracted from a given language model has higher predictive power for individuals of a certain cognitive profile.

Results. Figure 3 presents the relative gain or loss in predictive power of surprisal or entropy for individuals in distinct high- vs. low-performing subgroups with respect to a given score. The pat-

tern is consistent across surprisal and contextual entropy, and we can see a trend towards a negative PPG across all cognitive domains, indicating that the models' predictive power is higher for RTs of low-performing individuals. In particular for individuals with high scores in the verbal intelligence RIAS test, adding surprisal as a predictor does not improve reading time predictions as well as for individuals with lower scores in the same test, especially when extracted from models of the GPT-2 family. One of the few exceptions to the trend of a negative PPG (*i.e.,* a higher gain for low- over high-performing individuals) is the gain in predictive power for surprisal extracted from Mixtral when predicting RTs for high-performing individuals. Similarly, adding the scores for StrRTEffect, there is a positive PPG (higher gain in predictive power for high- over low-performing individuals), while for the scores of SimRTEffect, there is a negative PPG (higher gain in predictive power for low- over high-performing individuals).

5 Discussion

Most previous works on the predictive power of surprisal and entropy on reading times have been conducted on the group-level and assumed that potential fluctuations in effect sizes on the individual level do not interfere with the predictions under investigation. Our findings show that (1) incorporating information on individuals' cognitive capacities and allowing them to modulate surprisal and entropy effects can increase the predictive power of these predictors and that (2) a given language model has higher predictive power of reading times for individuals with a certain cognitive profile. For instance, the predictive power of surprisal extracted from GPT-2 degrades for individuals with higher verbal intelligence.

In experiment 2 (cf. §4.4), we find that different psycholinguistic subjects' predictability effects are modulated according to their cognitive capabilities, and that different LMs will predict higher/lower fluctuations from the group-level effects. In combination with our results from experiment 3 (cf. §4.5), which shows that predictive power in reading times is dependent on an individual's cognitive profile, this suggests that the choice of LM we estimate surprisal from for psycholinguistic studies is crucial for downstream analyses.

It would be ill-advised to choose an LM such as Mixtral for surprisal estimation of individuals with high working memory and GPT-2 *large* for a individuals with low working memory, since we have demonstrated that the predictive power of surprisal estimated from GPT-2 *large* is higher compared to Mixtral—suggesting that it is a better proxy of processing effort. However, when working on the group-level, it might be worthwhile to consider choosing a model that is less biased, or, in other words, whose predictability measures are equally as good of a proxy for individuals differing with respect to their capacity in a given cognitive domain. For instance, while both GPT-2's predictive power degrades for high-verbal-intelligence subjects, it is smaller for GPT-2 *large*. On the other hand, when working with data from a specific cognitive group (e.g., low-reading fluency), be it for psycholinguistic research or NLP applications, choosing an LM whose predictability measures have shown to be a better proxy of processing effort for the group at hand might be worthwhile and could potentially lead to more accurate results.

Moreover, while the scope of this work was solely on investigating whether certain LMs are better at predicting reading times of individuals with certain cognitive profiles, it might be worthwhile to extend the investigation on whether text produced by a given LM is biased towards being able to be 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 taylored to specific cognitive 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 predictability effects on human reading times is uniform across cognitive profiles. In this work, we have shown that indeed, LMs do show systematic biases towards certain cognitive profiles in the prediction of reading times. This finding is a crucial first step towards investigating whether it is worthwhile to consider information on individual cognitive capacities in language modelling.

Limitations

Splitting the subjects at the median of their scores for 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 cluster 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. 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.

Furthermore, contrary to previous work who set the reading times of words that were skipped to zero (Wilcox et al., 2023a,b; Pimentel et al., 2023), we exclude these words from our analysis. This is psycholinguistically sensible, as a word that is skipped and a word that was fixated, even if its first-pass or total fixation duration is very low, are qualitatively different. However, for the investigation of the predictive power of surprisal and contextual entropy, it might be prudent to include skipped words. Pimentel et al. (2023) found that word-skipping effects are only predicted by contextual entropy and not by surprisal. Thus including skipped words would not have changed the overall pattern of the results, but it might have led to a slightly higher predictive power of contextual entropy.

Ethics Statement

Working with human data requires careful ethical considerations. The *Individual Differences Corpus* (InDiCo; Haller et al., 2023) utilized for this study follows ethical standards and has been approved by the responsible ethics committees.

References

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.

Das Digitale Wörterbuch der deutschen Sprache. 2016. *Das Digitale Wörterbuch der deutschen Sprache (DWDS)*. Berlin-Brandenburg Academy of Science.

Andrea de Varda and Marco Marelli. 2022. The effects of surprisal across languages: Results from native and non-native reading. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, 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.

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.

Ulf Grenander. 1967. *Syntax-controlled probabilities*. Division of Applied Mathematics, Brown University.

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.

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.

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.

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 *Findings of the Association for Computational Linguistics: EMNLP 2023*, 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.

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 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.

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 Philip Levy. 2022. Large-scale evidence for logarithmic effects of word predictability on reading time.

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 *CogSci*, 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.

Charles S Wetherell. 1980. Probabilistic languages: A review and some open questions. *ACM Computing Surveys (CSUR)*, 12(4):361–379.

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. 2020a. 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*.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020b. On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, pages 1707–1713. Cognitive Science Society.

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*.

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:

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

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 Indico

Cognitive capacity		**Test/Task**	**Abbreviation**
Cognitive Control	V	Stroop	StrRTEffect
	NV	Simon	SimRTEffect
		FAIR-2	FAIRLPR
Working Memory	V	Sentence span	SSmean
		Operation span	OSmean
	NV	Memory updating	MUmean
		Spatial short-term memory	SSTMRelScore
Intelligence		MWT-B	MWTScore
	V	RIAS 1: Question answering	RIASVixPR
		RIAS 3: Sentence completion	
	NV	RIAS 2: Excluding mismatching images	RIASNixPR
		RIAS 4: Visual search of missing elements	
	C	RIAS 1-4	RIASGixPR
Reading Fluency	L	SLRT-II, word reading	SLRTWord
	NL	SLRT-II, pseudo-word reading	SLRTPseudo

Table 3: Psychometric tests conducted with all participants. *DV(s)* refers to dependent variables; *RT* refers to response latency, ACC to response accuracy, V to verbal, NV to non-verbal, C to combined, L to lexical, NL to non-lexical. Note that RIASGixPR refers to the average score obtained from RIAS 1-4.