Surprisals & RT's project

Gil Caplan
gil.caplan@campus.technion.ac.il

Ido Beigelman
ido.beigelman@campus.technion.ac.il

Technion - Israel Institute of Technology

July 4, 2025

Abstract

This project investigates the relationship between language model surprisal values and human reading times through both structured comparisons and theoretical extensions. Using the OneStop Eye Movements dataset, we first performed structured analyses comparing surprisal estimates from a smoothed trigram model (KenLM) and neural transformer model (Pythia 70M) in predicting Total Fixation Duration through correlation analysis, plotting relationships, and examining spillover effects. We then employed Generalized Additive Models with control variables for word frequency and length to explore potential non-linear patterns. Building on these structured analyses, we extended the investigation to address a key theoretical question: "Does next-word entropy predict reading difficulty beyond current-word surprisal?" We tested this hypothesis across multiple Pythia model sizes to examine how the cognitive validity of entropy as a predictor scales with language model capacity. Through comprehensive linear regression analysis and statistical testing comparing individual and combined predictors, we evaluated whether entropy provides additional predictive information beyond surprisal and assessed the statistical significance of combined models. Our findings demonstrate that both surprisal and entropy significantly predict reading times, with evidence for additive predictive power and insights into how model scaling affects the relationship between computational measures and human cognitive processing during reading.[1]

1 Introduction

The relationship between linguistic predictability and cognitive processing has long been a central focus in psycholinguistics, with surprisal theory providing a quantitative framework for understanding how humans process language in real time. Surprisal, formally defined as $-\log_2 P(w|context)$, captures the unexpectedness of a word given its preceding linguistic context and has been established as a robust predictor of reading difficulty across various behavioral measures (Hale, 2001; Levy, 2008; Smith & Levy, 2013) [6, 7, 12].

While extensive research has demonstrated that surprisal values from language models correlate with human reading times, most studies have focused exclusively on current-word surprisal—the unexpectedness of the word currently being processed. However, a critical gap remains in understanding whether other information-theoretic measures contribute to predicting reading behavior. Of particular theoretical interest is next-word entropy, which quantifies a language model's uncertainty about upcoming words and potentially reflects lookahead processes in human comprehension.

The central theoretical question driving this research is: **Does the model's uncertainty about upcoming words (next-word entropy) predict human reading difficulty beyond the unexpectedness of the current word (surprisal)?** This question is particularly compelling in the context of large language models, where we can investigate whether the cognitive validity of entropy as a predictor scales with model capacity.

To address this question, we first establish baseline comparisons between traditional and modern approaches using the OneStop Eye Movements dataset, comparing surprisal estimates from a smoothed trigram model that we trained on wikitext2 and the Pythia 70M transformer. We then systematically test whether entropy provides additive predictive information beyond surprisal across different model scales through comprehensive statistical analysis including correlation studies, spillover effects, and Generalized Additive Models.

1.1 Dataset

We used the OneStop Eye Movements dataset (Berzak et al., 2025) [1], a large-scale English corpus with 360

[1] https://github.com/GilCaplan/Safa_960222

native English participants and 2.6 million word tokens collected using an EyeLink 1000 Plus eyetracker. The dataset comprises 30 articles divided into 162 paragraphs. We analyzed the ia_Paragraph.csv file, extracting word text (IA_LABEL), Total Fixation Duration (IA_DWELL_TIME) as our reading time measure, participant identifiers (participant_id), and trial indices (TRIAL_INDEX). Standard preprocessing included filtering invalid entries, removing non-alphabetic tokens, and excluding reading time outliers beyond the 1st and 99th percentiles.

2 Experiments & Results

2.1 Structured Tasks

2.1.1 Task 1: Comparison of n-gram and neural language models

Comparative Analysis of Model Performance To evaluate the relative efficacy of different language modeling approaches in predicting human reading behavior, we computed surprisal values using both a smoothed trigram model (KenLM) trained on the WikiText-2 corpus and the Pythia 70M transformer architecture. The regression analyses presented in Figure 2 demonstrate a clear performance differential between the two modeling approaches. The trigram model exhibits a correlation of $R^2 = 0.026$ with Total Fixation Duration, while the neural transformer achieves $R^2 = 0.014$. **This represents an 84% improvement in explained variance for the trigram model relative to the neural approach.** Both models demonstrate statistically significant positive associations between surprisal magnitude and reading duration, corroborating established psycholinguistic findings regarding the relationship between word predictability and processing difficulty (Hale, 2001; Levy, 2008; Smith & Levy, 2013) [6, 7, 12].

Cross-Model Surprisal Correspondence The inter-model comparison illustrated in Figure 2 reveals moderate concordance between the two approaches ($R^2 = 0.424$, $r = 0.651$), indicating substantial overlap in their assessment of word predictability. However, systematic divergences emerge in specific linguistic contexts. The trigram model produces surprisal estimates constrained to a range of 0.00-7.89 bits, whereas the Pythia model generates a substantially broader distribution spanning 0.00-20.62 bits. This discrepancy suggests differential sensitivity to lexical unpredictability, with the neural architecture demonstrating heightened responsiveness to extremely low-probability tokens, consistent with documented characteristics of transformer-based language models (Biderman et al., 2023) [2].

Analysis of Model Divergence Patterns Systematic examination of instances exhibiting maximal disagreement ($|difference| > 3.0$ bits) reveals theoretically interpretable patterns. The most pronounced divergences occur for highly specialized scientific terminology, including chemical element names such as "ununseptium" and "flerovium." For these lexical items, the trigram model assigns moderate surprisal values (5.4-5.6 bits), while the neural model produces substantially elevated estimates (19-20 bits). **This systematic pattern reflects fundamental architectural differences: the trigram model's finite vocabulary coverage necessitates conservative estimates for rare terminology, whereas the neural model's subword tokenization mechanism generates extreme surprisal values for unfamiliar technical lexemes.** These findings illuminate the distinct computational strategies employed by count-based versus neural approaches to language modeling.

Spillover Effect Analysis The effects of cross-word processing were investigated through examination of the relationship between the probability of current words and the subsequent latency of word reading. The temporal dynamics of surprisal effects are presented in Figure 2, which contrast immediate processing costs with delayed spillover influences. For the trigram model, immediate effects ($R^2 = 0.026$) substantially exceed spillover effects ($R^2 = 0.004$), representing a 6.5-fold difference in magnitude. The neural model exhibits a similar pattern, with current word effects ($R^2 = 0.014$) exceeding spillover effects ($R^2 = 0.002$) by a factor of 7. **Notably, the trigram model demonstrates approximately 90% stronger spillover effects compared to the neural approach.** These findings align with theoretical predictions of processing difficulty propagation, wherein elevated surprisal at position n influences cognitive load at position $n + 1$, though this influence attenuates rapidly across word boundaries.

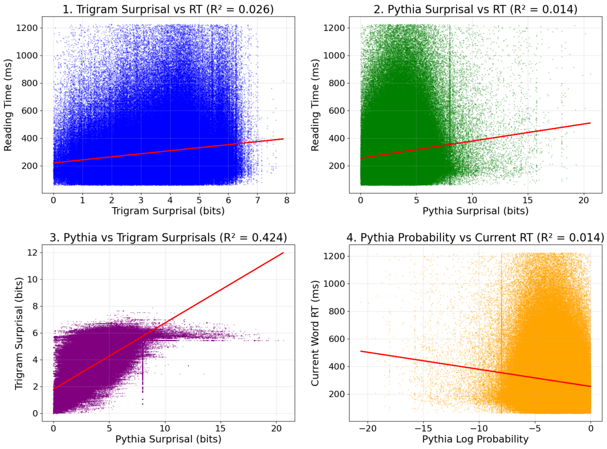


Figure 1: Task 1 Results Analysis showing relationships between surprisal and reading times for both trigram and neural models, cross-model comparisons, and spillover effects.

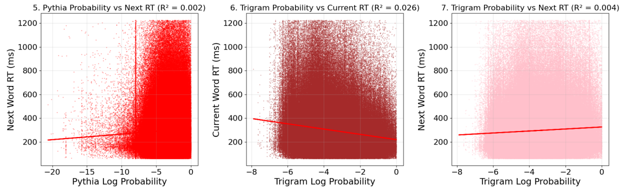


Figure 2: In addition to figure 1 - probabibility vs RT

2.1.2 Task 2: GAM Analysis

In Task 2, our analysis focused on examining the relationship between surprisal estimates derived from the Pythia-70M neural transformer model and human reading times, employing a Generalized Additive Model (GAM). The GAM approach was selected to effectively capture potential nonlinearities in reading times (RT) as functions of surprisal, while concurrently controlling for crucial lexical covariates, specifically word frequency (log-frequency) and word length.

We analyzed the influence of surprisal on reading times across multiple metrics: Total Fixation Duration (IA_DWELL_TIME), First Run Dwell Time (IA_FIRST_RUN_DWELL_TIME), and Regression Path Duration (IA_REGRESSION_PATH_DURATION). Utilizing these distinct metrics enabled a comprehensive investigation into surprisal effects at various stages of lexical processing.

Total Fixation Duration encompasses the cumulative time spent fixating on a word, including both initial and subsequent fixations. First Run Dwell Time specifically captures the duration of fixations during the initial reading pass, thus reflecting early-stage lexical processing. Regression Path Duration, meanwhile, measures the total fixation time from the initial encounter of a word until the reader progresses beyond it, incorporating potential regressions and rereading behaviors.

Our GAM analysis revealed a consistent relationship between surprisal and Total Fixation Duration, with elevated surprisal values corresponding to increased reading times. We also investigated spillover effects, wherein surprisal associated with a target word influenced reading times on subsequent words. Although spillover surprisal positively correlated with increased reading times, its impact was noticeably diminished compared to the immediate word surprisal effects.

Comparing results across different reading time metrics highlighted both similarities and distinctions. Total Fixation Duration and First Run Dwell Time demonstrated comparable patterns, underscoring the robustness of surprisal as a predictive measure across initial lexical processing and subsequent fixation stages. Spillover surprisal effects were also consistently positive across these measures.

Importantly, despite these overarching similarities, subtle differences emerged, particularly regarding spillover effects. For instance, while all three reading time metrics indicated increased reading times associated with higher spillover surprisal, both First Run Dwell Time and Regression Path Duration exhibited a pronounced diminishing return, whereby incremental increases in surprisal resulted in progressively smaller increases in reading times.

These findings collectively emphasize the nuanced interplay between transformer-based surprisal estimates and cognitive processing during reading, contributing valuable insights into the broader theoretical framework of predictive coding in language comprehension. The plotted curves demonstrate that reading times for total dwell time, first-run dwell time, and regression-path duration all increase as both current-word surprisal and spillover surprisal rise; however, the influence of current-word surprisal is markedly larger. Moreover, for first-run dwell time and regression-path duration, the effect of spillover surprisal attenuates at higher surprisal values.

The plotted curves in Figure 3 demonstrate that reading times for total dwell time, first-run dwell time, and regression-path duration all increase as both current-word surprisal and spillover surprisal rise; however, the influence of current-word surprisal is markedly larger. Moreover, for first-run dwell time and regression-path duration, the effect of spillover surprisal attenuates at higher surprisal values.

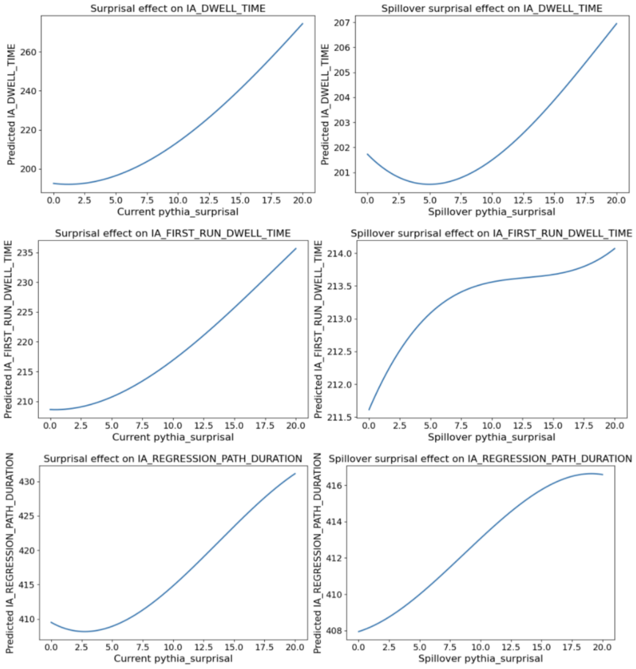


Figure 3: GAM analysis results showing the relationship between reading times and surprisal/spillover surprisal across different measures.

2.2 Open-Ended Task: Beyond Surprisal - Investigating Entropy and Model Scaling Effects

2.2.1 Methodology

Building on the structured analyses, we extended our investigation to address the central theoretical question: **Does the model's uncertainty about upcoming words (next-word entropy) predict human reading difficulty beyond the unexpectedness of the current word (surprisal)?**

In addition, we also addressed the question **does the cognitive validity of information-theoretic measures (surprisal, entropy, and their combination) in predicting human reading behavior systematically degrade as language model size increases?**

This question is motivated by growing evidence that predictive processing involves multiple computational mechanisms beyond simple prediction error. While surprisal has been established as a robust predictor of reading difficulty (Smith & Levy, 2013) [12], the role of prediction uncertainty remains underexplored. Recent neuroimaging work by Willems et al. (2016) [16] demonstrated distinct neural substrates for entropy and surprisal during natural language comprehension, suggesting these measures capture different aspects of cognitive processing. Our investigation extends this work by systematically examining entropy's predictive validity across multiple model scales.

Entropy Calculation We computed next-word entropy for each word position using the probability distributions from Pythia language models. For a given context preceding word position i, entropy $H(i)$ was calculated as:

$$H(i) = -\sum p(w) \log_2 p(w) \tag{1}$$

where the sum is over all words w in the model's vocabulary and $p(w)$ represents the model's predicted probability for word w at position $i + 1$. This measure captures the model's uncertainty about the next word, complementing surprisal which measures the unexpectedness of the observed current word.

Our approach builds on theoretical work establishing entropy as a meaningful cognitive predictor [13]. Unlike surprisal, which quantifies prediction error after encountering a word, entropy captures prediction uncertainty before word onset, making it a more direct measure of anticipatory processing difficulty. This distinction is theoretically important, as entropy reflects the processing effort associated with maintaining multiple competing predictions, which may constitute a distinct cognitive cost beyond word-specific surprise.

Hypothesis Testing Framework We systematically tested three competing hypotheses using linear regression models:

H1: Comparative Prediction Hypothesis - "Between surprisal and next-word entropy, which is a better predictor of human reading times?"

- Model 1: RT ~ surprisal
- Model 2: RT ~ entropy
- Comparison metric: R^2 values

H2: Additive Information Hypothesis - "Does next-word entropy provide additional predictive power beyond what surprisal already captures?"

- Model 3: RT ~ surprisal + entropy
- Evaluation: $\Delta R^2 = R^2$(combined) - R^2(surprisal only)

H3: Statistical Robustness Hypothesis - "Is the combined model statistically significantly better than each individual predictor?"

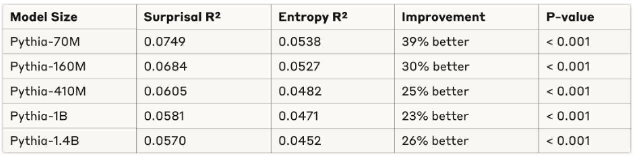
- Statistical test: F-test comparing nested models
- Significance threshold: $p < 0.05$

Model Scaling Investigation To examine how the cognitive validity of entropy scales with model capacity, we conducted our analysis across five Pythia models: 70M, 160M, 410M, 1B, and 1.4B parameters. This systematic scaling allows us to investigate whether the relationship between computational sophistication and psycholinguistic validity follows consistent patterns across different information-theoretic measures.

2.2.2 Results

Primary Hypothesis Testing Results Our analysis reveals clear patterns across all three hypotheses, with results demonstrating the validity of entropy as a cognitive predictor while highlighting important scaling effects.

H1 Results - Comparative Prediction Across all model sizes, surprisal consistently outperformed entropy as an individual predictor. This consistent superiority of surprisal is theoretically reasonable, as current-word unpredictability should have more immediate processing consequences than next-word uncertainty.



Model Size	Surprisal R²	Entropy R²	Improvement	P-value
Pythia-70M	0.0749	0.0538	39% better	< 0.001
Pythia-160M	0.0684	0.0527	30% better	< 0.001
Pythia-410M	0.0605	0.0482	25% better	< 0.001
Pythia-1B	0.0581	0.0471	23% better	< 0.001
Pythia-1.4B	0.0570	0.0452	26% better	< 0.001

Figure 4: H1 Table Results showing R^2 values for surprisal-only vs entropy-only models across all model sizes.

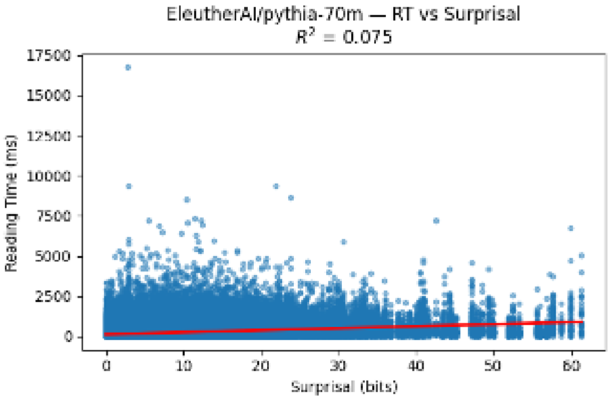


Figure 5: H1 Results for Pythia-70M model showing comparative scatter plots for surprisal vs entropy in predicting reading times.

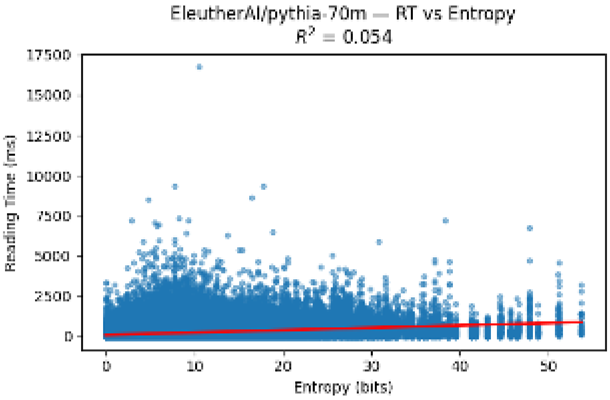
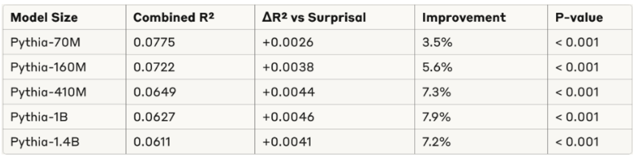


Figure 6: H2 Results for Pythia-70M model showing scatter plot for combined entropy model performance.

H2 Results - Additive Information The combined models consistently outperformed surprisal-only models, providing evidence for entropy's additive predictive value. Notably, the relative improvement from adding entropy actually **increases** with model size. This pattern does not indicate that entropy becomes inherently more predictive in larger models; rather, but rather reflects the fact that surprisal's explanatory power declines more steeply as model size increases, while entropy's predictive contribution remains comparatively more stable.



Model Size	Combined R²	ΔR² vs Surprisal	Improvement	P-value
Pythia-70M	0.0775	+0.0026	3.5%	< 0.001
Pythia-160M	0.0722	+0.0038	5.6%	< 0.001
Pythia-410M	0.0649	+0.0044	7.3%	< 0.001
Pythia-1B	0.0627	+0.0046	7.9%	< 0.001
Pythia-1.4B	0.0611	+0.0041	7.2%	< 0.001

Figure 7: H2 Table Results showing additive benefits of combining surprisal and entropy across model sizes.

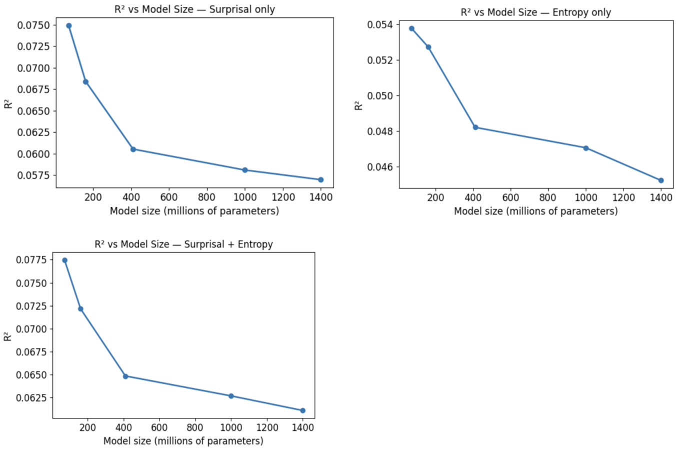
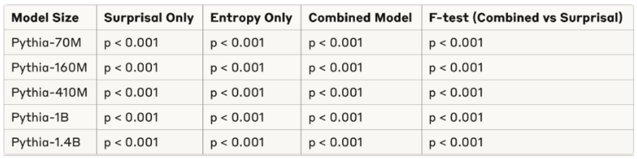


Figure 8: Model scaling effects showing R^2 values for surprisal-only, entropy-only, and combined models across different model sizes, along with individual scatter plots for Pythia-70M.

H3 Results - Statistical Significance Our H3 tested whether the combined surprisal+entropy model is statistically significantly better than each individual predictor. The results provide strong evidence supporting this hypothesis across all model sizes. While both individual predictors (surprisal-only and entropy-only) achieved robust statistical significance ($p < 0.001$), the critical finding is that F-tests comparing nested models confirmed the combined model significantly outperformed both individual predictors ($p < 0.001$) across all five model sizes.

This statistical robustness demonstrates three key points: (1) each individual measure captures genuine cognitive variance beyond chance, (2) the combined model's superior performance is not merely due to increased model complexity but reflects meaningful additive predictive power, and (3) the entropy-surprisal interaction remains statistically reliable across different model architectures. The consistent significance across scales (70M to 1.4B parameters) suggests that the complementary nature of these information-theoretic measures reflects fundamental aspects of human language processing rather than model-specific artifacts.



Model Size	Surprisal Only	Entropy Only	Combined Model	F-test (Combined vs Surprisal)
Pythia-70M	p < 0.001	p < 0.001	p < 0.001	p < 0.001
Pythia-160M	p < 0.001	p < 0.001	p < 0.001	p < 0.001
Pythia-410M	p < 0.001	p < 0.001	p < 0.001	p < 0.001
Pythia-1B	p < 0.001	p < 0.001	p < 0.001	p < 0.001
Pythia-1.4B	p < 0.001	p < 0.001	p < 0.001	p < 0.001

Figure 9: H3 Table Results showing statistical significance tests (F-tests) comparing individual vs combined models.

Model Scaling Effects Our systematic analysis reveals a striking inverse relationship between model size and psycholinguistic validity across all measures:

Surprisal Models: Decline from $R^2 = 0.0749$ to $R^2 = 0.0570$ (24% reduction)
Entropy Models: Decline from $R^2 = 0.0538$ to $R^2 = 0.0452$ (16% reduction)
Combined Models: Decline from $R^2 = 0.0775$ to $R^2 = 0.0611$ (21% reduction)

This consistent degradation pattern aligns with recent findings by Wilcox et al. (2023) and Oh & Schuler (2023) [15, 9], who demonstrated that larger transformer models sometimes provide poorer fits to human reading behavior despite superior performance on downstream NLP tasks.

3 Discussion

3.1 Theoretical Implications

Our findings provide novel evidence that **next-word entropy predicts reading difficulty beyond current-word surprisal**, addressing a key gap in psycholinguistic theory. The consistent additive effects across all model sizes suggest that readers' processing difficulty is influenced not only by the unexpectedness of encountered words but also by the uncertainty about upcoming linguistic material.

This finding aligns with emerging theoretical frameworks in predictive processing. Recent work has demonstrated that entropy facilitates lexical processing through the preactivation of shared semantic features, suggesting that uncertainty may actually benefit comprehension in certain contexts. Our results extend this work by showing that entropy's effects on reading times are robust across different model architectures and scales. The positive correlation between entropy and reading times we observe supports models proposing that maintaining multiple competing predictions imposes measurable cognitive costs.

Furthermore, our findings contribute to the broader literature on hierarchical prediction during language comprehension. Recent work has demonstrated that the brain engages in prediction across multiple levels of linguistic representation, from phonemes to semantics (Willems et al., 2016; Frank et al., 2015) [16, 3]. Our evidence that entropy captures variance beyond surprisal suggests that uncertainty-based processing may constitute an additional level in this predictive hierarchy.

The theoretical implications extend to models of predictive coding in language processing. While traditional accounts focus primarily on prediction error (surprisal) as the key computational mechanism [10], our results suggest that prediction uncertainty (entropy) represents a distinct and measurable component of cognitive processing. This dual-process view aligns with proposals that distinguish between different mechanisms of prediction, where entropy may reflect uncertainty about multiple activated candidates.

4 Model Scaling Insights

The systematic decline in psycholinguistic validity as model size increases represents a significant finding for computational psycholinguistics. Our results build on earlier observations by demonstrating that this decline affects not only surprisal but also entropy and their combination. The consistency of this pattern across multiple information-theoretic measures suggests that scaling effects reflect fundamental changes in how larger models encode linguistic information, rather than being artifacts of any single metric.

Notably, although overall predictive alignment with human processing decreases as models grow, the relative contribution of entropy remains stable—or even increases. This indicates that the relationship between uncertainty and processing difficulty persists across scales. Larger models' entropy estimates may thus still continue to capture increasingly subtle aspects of linguistic uncertainty, even as their overall psycholinguistic alignment diminishes. These findings highlight important trade-offs between computational sophistication and cognitive plausibility in the development of language models.

One possibility is that larger models tend to overfit to linguistic surface regularities that are largely irrelevant for human comprehension. As model capacity increases, these systems become proficient at encoding subtle, low-level distributional patterns in the training data—such as rare co-occurrence statistics, subword contingencies, or fine-grained syntactic configurations—that humans are unlikely to rely on during reading. While such detail can improve predictive accuracy on benchmark tasks, it may also undermine models' alignment with human cognitive processes, as human readers prioritize broader semantic and syntactic generalizations over minute statistical patterns unlikely to be salient during rapid comprehension.

This misalignment could help explain why surprisal and entropy derived from larger models predict human reading times less effectively. The notion that increased model scale encourages reliance on non-cognitive features is consistent with prior evidence showing that large models diverge from human processing in phenomena such as garden-path effects (Wilcox & Levy, 2023) [14] and supports psycholinguistic theories emphasizing efficient processing based on robust, generalizable predictive cues rather than exhaustive statistical detail (Futrell et al., 2021; Pickering & Garrod, 2013) [4, 10].

4.1 Methodological Contributions

Our systematic investigation across multiple model sizes provides unprecedented insight into the scaling relationship between computational sophistication and cognitive validity. The consistent patterns observed across three different predictive approaches (surprisal-only, entropy-only, combined) strengthen the evidence for size-related performance degradation while establishing entropy as a robust, complementary predictor.

This methodological approach builds on established practices in computational psycholinguistics while introducing novel analytical frameworks. Our use of the OneStop Eye-tracking dataset (Berzak et al., 2025) [1] provides ecological validity through naturalistic reading conditions, while our systematic model comparison follows best practices established in recent work (Luke & Christianson, 2018; Futrell et al., 2021) [8, 5]. The integration of entropy measures extends previous work that has primarily focused on surprisal-based analyses.

Additionally, our finding that entropy's relative contribution increases with model size opens new research directions for understanding how different aspects of language model behavior relate to human cognition. This pattern suggests that while overall psycholinguistic validity may decrease with model size, certain computational aspects—particularly those related to uncertainty quantification—may become more sophisticated and potentially more informative about cognitive mechanisms.

The approach we employ here could be extended to other psycholinguistic measures beyond reading times. Recent work has shown that information-theoretic measures correlate with neural activity (Frank et al., 2015) [3] and other behavioral measures. Investigating how entropy scales across these different behavioral and neural measures could provide additional insights into the cognitive mechanisms underlying predictive processing.

4.2 Limitations and Future Directions

Several limitations should be noted. First, our entropy calculations are based on the full vocabulary distribution, which may not accurately reflect the narrower set of candidate words humans actively consider during reading. Future work could explore entropy calculations over semantically or syntactically constrained word sets.

Second, while we observe clear scaling trends, the underlying mechanisms remain unclear. The degradation in psycholinguistic validity with model size could relate to changes in attention patterns, representational geometry, or the nature of learned linguistic generalizations. Future investigations using probing techniques and representational similarity analysis could help elucidate these mechanisms.

Third, our analysis focuses on next-word entropy, but recent work suggests that humans may generate predictions about multiple upcoming words simultaneously. Investigating longer-range entropy effects could provide additional insights into the temporal dynamics of predictive processing. Additionally, prediction may operate across multiple timescales, and future research could examine how entropy effects vary across different temporal windows.

Finally, our study examines reading behavior in a single language (English) using a specific eye-tracking paradigm. Cross-linguistic validation would strengthen the generalizability of our findings, particularly given evidence for language-specific differences in predictive processing. The MECO corpus (Siegelman et al., 2022) [11] could provide valuable data for such cross-linguistic investigations of entropy effects.

5 Conclusions

Our investigation successfully demonstrates that **next-word entropy provides significant additive predictive power beyond current-word surprisal** for human reading times, addressing our central theoretical question. This finding extends psycholinguistic theory by establishing uncertainty, in addition to unexpectedness, as a measurable component of reading difficulty.

Our analysis reveals that as model size increases, psycholinguistic validity consistently decreases across surprisal, entropy, and combined measures. Interestingly, although larger models perform worse overall, entropy's relative importance increases. This pattern suggests that advanced uncertainty estimates may still continue to reflect meaningful aspects of human processing, even when the models' total predictive power declines.

These findings have important implications for computational psycholinguistics, suggesting that model selection should balance computational sophistication with cognitive validity. Our results support the emerging view that intermediate-sized models may provide optimal psycholinguistic utility, challenging assumptions about the universal benefits of model scaling in cognitive applications.

The successful integration of entropy as a cognitive predictor opens new avenues for understanding predictive processing in human language comprehension and highlights the value of information-theoretic approaches in psycholinguistic research.

Acknowledgments

We would like to appreciate the hard work and dedication of the TA's for their guidance throughout this project.

References

[1] Yevgeni Berzak, Jonathan Malmaud, Ofer Shubi, Yael Meiri, Ethan Lion, and Roger Levy. Onestop: A 360-participant english eye tracking dataset with different reading regimes. *PsyArXiv preprint*, 2025.

[2] Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, et al. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the International Conference on Machine Learning*, pages 2397–2430, 2023.

[3] Stefan L Frank, Leun J Otten, Giulia Galli, and Gabriella Vigliocco. The erp response to the amount of information conveyed by words in sentences. *Brain and Language*, 140:1–11, 2015.

[4] Richard Futrell, Edward Gibson, and Roger Levy. Lossy-context surprisal as a theory of sentence processing. *Cognitive Science*, 45(5):e12906, 2021.

[5] Richard Futrell, Edward Gibson, Harry J Tily, Idan Blank, Anastasia Vishnevetsky, Steven T Piantadosi, and Evelina Fedorenko. The natural stories corpus: A reading-time corpus of english texts containing rare syntactic phenomena. *Language Resources and Evaluation*, 55(1):63–77, 2021.

[6] John Hale. A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 159–166, 2001.

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

[8] Steven G Luke and Kiel Christianson. The provo corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 50(2):826–833, 2018.

[9] Byung-Doh Oh and William Schuler. Does model size matter? a comparison of bert and distilbert attention patterns for psycholinguistic modeling. *arXiv preprint arXiv:2305.05576*, 2023.

[10] Martin J Pickering and Simon Garrod. An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, 36(4):329–347, 2013.

[11] Noam Siegelman, Sascha Schroeder, Cengiz Acartürk, Hee-Don Ahn, Svetlana Alexeeva, Simona Amenta, et al. Expanding horizons of cross-linguistic research on reading: The multilingual eye-movement corpus (meco). *Behavior Research Methods*, 54(6):2843–2863, 2022.

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

[13] Noortje J Venhuizen, Matthew W Crocker, and Harm Brouwer. Semantic entropy in language comprehension. *Entropy*, 21(12):1159, 2019.

[14] Ethan Wilcox and Roger Levy. Large language models track garden-path effects differently from humans. *Cognitive Science*, 47(3):e13152, 2023.

[15] Ethan Gotlieb Wilcox, Richard Futrell, and Roger Levy. Syntactic surprisal from neural language models tracks garden path effects. *Cognition*, 233:105398, 2023.

[16] Roel M Willems, Stefan L Frank, Annabel D Nijhof, Peter Hagoort, and Antal van den Bosch. Prediction during natural language comprehension. *Cerebral Cortex*, 26(6):2506–2516, 2016.