Journal of Memory and Language 146 (2026) 104677

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Language models that match reader experience are better predictors of reading times☆

Iza Škrjanec [a,b,*], Vera Demberg [a]

[a] Saarland University, Campus, Saarbrücken, 66123, Germany
[b] Zuse School ELIZA, Hochschulstr. 10, Darmstadt, 64289, Germany

ARTICLE INFO

Dataset link: https://osf.io/78jdq/

Keywords:
Language models
Surprisal
Domain expertise
Reading times

ABSTRACT

Humans differ in the language experience that they accumulate, due to differing interests, reading habits and profession. This experience can be expected to affect their linguistic expectations when reading texts from domains that are very familiar to them. The present article explores whether language models trained to match the experience of readers produce surprisal estimates that more accurately predict the reading times of those readers than the usually employed general language models. We use a German eye-tracking corpus of biology and physics students reading expository texts from these domains. We adapt a neural language model to the experience of these two groups of readers via two domain adaptation methods and varying amounts of training data. The evaluation against one early and two late reading measures suggests that aligning language models with the readers' experience to predict the processing effort results in a better fit on late measures than using a model with a high linguistic accuracy. Our findings highlight the opportunities for exploring the cognitive plausibility of language models with respect to psychological constructs.

Introduction

Surprisal, as defined by Hale (2001) and Levy (2008), quantifies the predictability of a linguistic unit, such as a word, within a given context. It has emerged as a key construct in psycholinguistics for modeling human processing effort during language comprehension. Numerous studies have linked surprisal to behavioral measures like reading times (Demberg & Keller, 2008; Smith & Levy, 2013; Wilcox et al., 2020) and neural responses such as the N400 (Frank et al., 2015; Michaelov et al., 2024; Szewczyk & Federmeier, 2022).

Formalized as the negative logarithm of the probability of a word given context, surprisal is commonly derived from language models (LMs, Hale, 2001). Traditional psycholinguistic studies have often conflated two views of surprisal estimation: one that reflects the linguistic structure and distribution of a text as captured by the training data of a language model, and another that aligns more closely with human readers' cognitive processing mechanisms. While early studies emphasized the former—assuming that models with a higher linguistic accuracy and thus lower perplexity would inherently align better with human reading behavior (Frank & Bod, 2011; Goodkind & Bicknell, 2018)—recent findings challenge this assumption. Contemporary large-scale language models, despite achieving remarkably low perplexity, often perform poorly in predicting human correlates like reading times (de Varda & Marelli, 2023; Oh & Schuler, 2023a, 2023b; Shain et al., 2024). This calls into question whether the ability to accurately predict upcoming words alone suffices for cognitive plausibility.

In light of these recent findings on contemporary language models, our study seeks to address this challenge by investigating the effects of aligning language models with the background knowledge and experience of individual readers. Humans differ in linguistic expectations due to variations in reading habits, interests, and professional expertise. These differences, particularly in domain-specific contexts, are likely to influence processing effort. For example, previous studies show that readers with greater domain knowledge process text more efficiently, exhibiting shorter reading times and fewer regressions (Jian & Ko, 2014), while contradictions in information in the text and prior knowledge trigger more regressions (van Moort et al., 2020). However, existing language models fail to capture such individual variability, as they typically use a one-size-fits-all approach to surprisal estimation.

To explore this, we focus on reader-aligned and text-aligned surprisal in predicting reading times. Reader-aligned surprisal incorporates a model adapted to the reader's domain expertise, while text-aligned surprisal estimates processing effort based solely on the linguistic properties of the text. By comparing these approaches, we aim to assess how aligning language models with human experience can improve

☆ This article is part of a Special issue entitled: 'Language Models & Psycholinguistics' published in Journal of Memory and Language.

* Corresponding author at: Saarland University, Campus, Saarbrücken, 66123, Germany.
E-mail addresses: skrjanec@lst.uni-saarland.de (I. Škrjanec), vera@lst.uni-saarland.de (V. Demberg).

https://doi.org/10.1016/j.jml.2025.104677
Received 23 October 2023; Received in revised form 12 July 2025; Accepted 20 July 2025
Available online 3 October 2025

their explanatory power for reading behavior. This complements recent advances in language model adaptation, where techniques such as fine-tuning and adapter training have been employed to improve performance in specialized domains (e.g., Gururangan et al., 2020; Welch et al., 2022). Unlike prior adaptation studies, which focus on text alignment, we emphasize aligning models to readers, considering their domain-specific linguistic expectations and how these influence reading effort.

Our work builds on previous analyses of the Potsdam Textbook Corpus (Jakobi et al., 2025), a German eye-tracking dataset that records university students' reading behavior for physics and biology texts. This corpus provides a unique opportunity to study how domain expertise affects reading behavior, with expertise operationalized as biology students reading biology texts and physics students reading physics texts. Analyses show that this domain expertise significantly affects reading times, namely higher domain familiarity leads to faster reading (Jakobi et al., 2025; Škrjanec et al., 2023).

To better understand how domain expertise and linguistic predictability influence reading behavior, it is essential to consider how these factors affect early and late reading measures. Early measures, such as first-pass reading time, predominantly capture pre-lexical and lexical processing stages, including word recognition and lexical access. Late measures, such as go-past time and total reading time, additionally reflect higher-level cognitive processes, including integration with the previous context (Radach & Kennedy, 2013; Rayner & Liversedge, 2011). Prior research demonstrates that predictability influences early measures (Staub, 2015), but large-scale analyses of naturalistic reading data have consistently shown robust effects on late measures as well (Shain et al., 2024; Wilcox, Pimentel et al., 2023). Domain expertise also plays a critical role in both types of measures. In early measures, expertise facilitates the activation of relevant concepts and anticipation of upcoming words, enabling faster word recognition and lexical access. In late measures, expertise aids in sentence integration and the retention of information, further reducing cognitive effort during reading (Jian & Ko, 2014). By analyzing both early (first-pass time) and late reading measures (go-past and total time), we aim to capture the nuanced ways in which linguistic predictability and domain knowledge interact, shedding light on their distinct and overlapping contributions to processing effort.

We formulate the following questions and address them in separate studies:

1. **Study 1**: Does experience in terms of domain familiarity affect reading, focusing on the measures of first-pass duration, go-past, and total reading time?
2. **Study 2**: Predictive power of reader-aligned surprisal: what is the effect of the amount of training data and domain adaptation method?
3. **Study 3**: Which alignment strategy (reader-aligned or text-aligned surprisal) is most beneficial for reading time prediction?

To answer these questions, we compare surprisal estimates from two domain adaptation methods: full fine-tuning, which updates all model parameters to specialize in a target domain, and adapters (Houlsby et al., 2019), which enable parameter-efficient adaptation by learning transformations of pre-trained representations. These methods allow us to systematically examine how varying degrees of specialization influence the fit of surprisal estimates to reading times across different measures of processing effort.

An important dimension of our study lies in exploring the parallels between human language comprehension and the predictive behavior of reader-aligned language models. Specifically, we examine human reading times and compare them to the word probability distributions generated by language models. Our results demonstrate that aligning language models with readers' domain expertise leads to more accurate surprisal estimates and a better fit to late measures of human reading times. However, we view these findings as only the starting point. Surprisal Theory as such (Hale, 2001; Levy, 2008) is agnostic to the actual mechanism that generated the conditional probability distribution over the vocabulary. Surprisal estimates are derived from the final layer of a language model, reflecting its output rather than the internal mechanisms that produce these predictions. While adapted language models exhibit prediction patterns that align more closely with human reading behavior, future research must go beyond behavioral similarity. A deeper understanding of the mechanisms driving these predictions, both the cognitive processes in humans and the internal computations of neural language models, is essential. Although a detailed exploration of mechanistic interpretability is beyond the scope of this study, we recognize its significance and return to it in the general discussion.

Background

The role of domain knowledge in language processing

Background knowledge plays a crucial role in shaping behavior and understanding during language processing. Numerous empirical studies report the effect of background knowledge on comprehension and show that sufficient knowledge of the text's topic facilitates correct inference-making and a higher recall, while a lack of required knowledge may result in comprehension difficulties (Kaakinen et al., 2003; Kendeou & van den Broek, 2007; Kendeou et al., 2004; Ozuru et al., 2009; Tarchi, 2010).

Processing difficulties due to limited lexical knowledge manifest in online measures like reading behavior. For instance, encountering individual unfamiliar or infrequent words leads to slower reading and increased revisits (Just & Carpenter, 1980; Lowell & Morris, 2014). This holds for familiarity with the vocabulary of specific domains as well. Scientific texts frequently incorporate specialized and less common vocabulary, requiring the understanding of domain-specific concepts and relations between them.

Jian and Ko (2014) report that readers with a higher level of background knowledge in physics had shorter first-pass reading times and lower regression rates and spent less time rereading in comparison to lower-knowledge readers. Additionally, encountering information that is inconsistent with the domain knowledge of the reader shows effects on neural and behavioral measures, eliciting a greater N400 amplitude (Troyer & Kutas, 2018; Troyer et al., 2020) or longer reading with more regressions (van Moort et al., 2020).

Despite its importance, integrating world or domain knowledge with linguistic input in models of comprehension remains a challenging task. Frank et al. (2008) highlight that many theoretical models of discourse comprehension struggle to account for the necessary world knowledge, leading to problematic predictions. Addressing this gap, Venhuizen et al. (2019) proposed a model based on simulations in a micro-world with limited vocabulary and knowledge. Their findings show that comprehension is better modeled when surprisal is informed by the interaction of lexical material and world knowledge, rather than relying on either text-based or knowledge-based surprisal alone.

Language model pre-training and domain adaptation

Recent work on modeling word-by-word processing during reading has typically used pre-trained Transformer (Vaswani et al., 2017) language models. While different language model architectures exist, we focus here on decoder-only LMs that are trained to generate text (next-word prediction objective) and consider only the preceding context.

In practice, language models typically undergo two stages of training: (1) the pre-training stage, and (2) transfer learning for the target domain/task. In the pre-training stage, the model parameters are updated based on the prediction error given the input training data. Modern language models are trained on enormous training corpora

that exceed the size of developmentally plausible data by multiple magnitudes (Warstadt & Bowman, 2022). The necessity of such vast amounts of training data is questionable: Zhang et al. (2021) report that basic syntactic and semantic features are successfully encoded with 10 million to 100 million words, while the remaining data exposure serves for learning higher-level skills like commonsense reasoning.

If the feature distribution of the target domain (i.e. the domain of the downstream task) differs from that of the pre-trained model, the model undergoes a second stage of training, in which it is adapted to the target domain or task via transfer learning (see Zhuang et al., 2021 for an overview). The idea is to transfer the readily learned linguistic representations and specialize the model for the downstream task, which usually results in higher performance (Gururangan et al., 2020).

Among the methods of transfer learning, we consider two that focus on domain adaptation: full fine-tuning and adaptation via adapter weights. The two techniques differ with respect to the base pre-trained model. The full fine-tuning paradigm optimizes all the weights of the pre-trained model based on the in-domain training data. This means that previously learned knowledge is fully updated to represent the target domain. In contrast, adapters offer a parameter-efficient method for transferring the learned features and learning the new domain. In our study, we use the so-called bottleneck adapters (Houlsby et al., 2019). In the bottleneck adapter approach, new layers (adapter weights, which consist of feed-forward layers) are inserted between existing layers of the pre-trained model. These weights are of smaller dimensions than the pre-trained model. During training, the layers of the pre-trained model are frozen and remain intact, while the weights of the adapters are updated with respect to the loss function on the training data. The role of the adapter weights is to provide transformations of frozen pre-trained weights given the target domain. With ever larger pre-trained LM, full fine-tuning is associated with high computational costs, while adapters achieve a similar performance (Pfeiffer et al., 2020) with fewer resources.

In terms of cognitive plausibility, the two adaptation methods offer distinct perspectives. Full fine-tuning fully adapts the model to the target domain, potentially at the cost of forgetting general linguistic features (Thompson et al., 2019). In contrast, adapters may more closely resemble how humans adjust their linguistic expectations based on the context (Dubey et al., 2006; Fine et al., 2013; Nieuwland & van Berkum, 2006; van Schijndel & Linzen, 2018), while preserving their general linguistic abilities. However, a key difference between both adaptation techniques and human language learning is that the model's lexicon remains unchanged during adaptation, which contrasts with human language processing (Gaskell & Ellis, 2009; Laine et al., 2014).

The motivation for exploring these two domain adaptation techniques lies in their potential trade-offs. On the one hand, full fine-tuning could be at a disadvantage due to overfitting to domain-specific properties at the expense of general linguistic features. On the other hand, it might better capture the characteristics of the target domain precisely because previously learned language properties are overwritten. Their performance in terms of perplexity is expected to be similar (Pfeiffer et al., 2020), but it remains unclear how their surprisal estimates will differ in predictive power when modeling human reading times.

Surprisal and its predictive power

Surprisal theory (Hale, 2001; Levy, 2008) posits that the human processing effort of a word is proportional to surprisal, which is based on the probability of a word given its preceding context (Eq. (1)). While surprisal theory is not specifically designed to predict eye movements during reading, it has been shown to correlate with reading measures from self-paced reading and eye-tracking corpora in different languages (e.g. by Boston et al., 2008; Demberg & Keller, 2008; Oh et al., 2022; Smith & Levy, 2013; Wilcox, Pimentel et al., 2023), as well as event-related potentials during language processing (e.g. in Michaelov et al., 2024).

$$surprisal(w_i) = -\log_2 p(w_i|w_1 \ldots w_{i-1}) \quad (1)$$

Surprisal has been studied for its predictive power for behavioral data from naturalistic reading (Demberg & Keller, 2008; Shain et al., 2024; Wilcox, Pimentel et al., 2023) and its sensitivity to specific phenomena, such as syntactic ambiguity (Huang et al., 2024; van Schijndel & Linzen, 2020), embedded structures (Hahn et al., 2022), semantic relatedness (Cong et al., 2023), plausibility and expectancy (Krieger et al., 2025; Michaelov et al., 2024).

The conditional probability term in surprisal is typically estimated using a language model (LM). With the advances of machine learning and natural language processing, surprisal is often estimated with large neural language models. Active research on LM quality, architecture and size shows that in comparison to n-gram models, probabilistic context-free grammar or neural Long Short-Term Memory language models (Aurnhammer & Frank, 2018; Fossum & Levy, 2012; Frank & Bod, 2011; Goodkind & Bicknell, 2018; Wilcox et al., 2020), neural Transformer language models provide surprisal estimates that lie closest to measured human processing effort (Merkx & Frank, 2021; Shain et al., 2024). While increasingly large Transformer LMs (such as the Generative Pre-trained Transformer, GPT, Radford et al., 2019) achieve lower and lower perplexities, indicating a higher linguistic accuracy, recent studies show a possible divergence between the linguistic quality and predictive power of human processing effort. Surprisal from LMs with either an extremely large number of parameters or training samples provides a worse fit than smaller GPT models (Oh & Schuler, 2023a, 2023b), but see Wilcox, Meister et al. (2023) for comparison.

Surprisal is an estimate of human processing effort, but an often overlooked aspect of this work is how well these estimates predict behavior across different readers or reader groups. In light of individual variation between readers, we can expect that modeling online measures can benefit from surprisal estimates informed by individual differences. A notable example is the analyses of surprisal effects for readers who speak English as a second language (Berzak & Levy, 2023; de Varda & Marelli, 2022), but they do not explore how to estimate incremental processing cost for L1 and L2 readers separately. Another study (Haller et al., 2024) investigated the predictive power of surprisal from large language models jointly with reading time data and results from psychometric tests. They found that surprisal estimates were more accurate for readers with lower performance on psychometric tests and that high performance readers might be less sensitive to predictability. They did not explore methods for adapting language models to represent specific reader profiles.

The effect of surprisal on different reading time measures

In this work, we focus on three measures: first-pass, go-past and total reading time (RT). First-pass RT (also called gaze duration) is an early measure summing all fixations on a word in the first pass before leaving the region. The measures of go-past and total RT are late measures. Go-past time of a word (also called regression-path duration) includes first-pass time as well as any regressive fixations after leaving the word to the left, until the point where reading progresses further to the right of the word. Total fixation time sums up the durations of all fixations on a word.

While it should be noted that there is no direct correspondence between the different steps of language processing and various reading time measures, research on eye movement during reading generally shares the view that early stages of processing are captured in early measures, and higher-level processing is reflected in late measures (Rayner & Liversedge, 2011). During the first-pass reading lexical access is said to take place. Go-past time indexes additional processing cost triggered by the word via regressions. Post-lexical semantic and

discourse integration will affect later measures, such as total reading time (Radach & Kennedy, 2013; Rayner & Liversedge, 2011).

The three chosen measures are contingent in that they include each other: first-pass is included in go-past time as well as in total reading time. Go-past time contains rereading of previous context during the first pass, thus including regressions triggered by the current word. As such, the measures often correlate (see e.g. de Varda et al., 2023).[1]

Previous research has also reported findings on the effects of surprisal on early and late measures, often resulting in smaller effects for first-fixation duration and typically find larger surprisal effects in first-pass times or later measures such as go-past, scan path and total reading time (de Varda & Marelli, 2023; Greenberg, 2023; Shain et al., 2024). For instance, Greenberg (2023) compares the amount of variance that surprisal can explain for different reading time measures and finds that the proportion of explained variance is largest in the latest reading time measures such as total reading times.

Data

The research reported in this article draws on two specialized corpora: the PoTeC corpus is a German eye-tracking corpus containing physics and biology texts. This corpus is used in our work to compare the surprisal estimates against the measured reading times. We furthermore collected domain-specific German physics and biology texts for domain adaptation of the pre-trained language models.

Eye-tracking dataset: Potsdam Textbook Corpus (PoTeC)

We evaluate how well the language models match the readers in the Potsdam Textbook Corpus (PoTeC, Jakobi et al., 2025), which includes eye-tracking measures gathered from 75 participants while reading six expository texts from biology and six texts from physics domain. All participants were German native speakers studying either physics (N = 32) or biology (N = 43) at university. For the experiment, they were asked to read each text for comprehension and answer text-related questions, as well as questions assessing their background knowledge about the topic of the text in general (independent of the text content).

Each text was taken from a textbook, has about 158 words (minimally 126 and maximally 180) and fits on a single screen. The corpus also contains word-level manual annotations for terminology with three labels: common words (level 0), generally known (level 1), and domain-specific technical terms (level 2). In total, there are 954 words in biology texts, 79% of these are of level 0, 11% are level 1, and 9% are technical terms of level 2. The distribution is similar in physics texts with 941 words in total, out of which 78% are of level 0, 15% of level 1, and 7% of level 2.

In total, PoTeC contains 142,125 data points (75 participants × (954 + 941) words). We exclude all data points that come from either the first or the last word in the text. Further, we exclude all data points that lie three standard deviations below or above a participant's average reading time. Data points with zero duration were also removed. We perform these steps for each of the three reading measures of interest: first-pass reading time (FPRT), go-past (GP) and total fixation time (TFT). For FPRT data, this results in 1350 data points (min. 934, max. 1622) per reader on average. For GP, in 1346 data points per (min. 935, max. 1623) reader on average. For TFT, this results in 1680 data points (min. 1381, max. 1829) per reader.

Fig. 1 presents an example from a biology text with the reading times for each group. We can see that the readers with a physics background read the text more slowly, as evidenced in both early effects in the first pass, and even larger effects in late measures.

In our study, we distinguish between experts and novices of the domain, i.e. biology students are experts in the biology domain, but termed "novices" when reading physics. Similarly, students majoring in physics are novices in biology, but experts in their academic domain. This distinction is an oversimplified operationalization of background knowledge, i.e. a physics student can be familiar with concepts from biology and is by no means an actual novice to the discipline. We also acknowledge the similarity between the disciplines of biology and physics: oftentimes undergraduate programs in one include courses from the other. Our distinction between "experts" and "novices" targets the level of expertise: we assume students majoring in physics are more familiar with the conceptual knowledge and vocabulary of the physics domain than the biology domain. This assumption is indeed warranted by the offline comprehension measures of the PoTeC dataset. Participants were given text-related comprehension questions for each text, as well as domain knowledge question on topics related to the texts, but not based on the text content. On average, students demonstrated greater background knowledge and more precise text comprehension when reading texts related to their field of study (see Table 2 in Škrjanec et al., 2023 for significance testing). This result also suggests that slower reading and rereading could not entirely make up for the lack of domain knowledge.

Comparing experts and non-experts within related disciplines has the drawback of excluding complete novices in the domain. However, selecting two entirely unrelated fields introduces the risk of confounding factors that could influence reading, such as differences in the general amount of reading required (e.g., participants majoring in literature versus those majoring in computer science). For instance, Troyer et al. (2023) encountered challenges in disentangling general print exposure from background knowledge when studying knowledge of the fictional world of Harry Potter. This difficulty likely arises because high-knowledge participants tend to read more overall, resulting in greater print exposure compared to low-knowledge readers.

Training data for the adaptation to the physics and biology domains

We use the German domain-general GPT-2 language model (GerPT2, Minixhofer, 2020) and adapt it to the biology and physics domains using domain-specific corpora as training data. To the best of our knowledge, there is no available training set in German for the two domains. Therefore, we create our own by scraping articles from two online sources: Wikipedia and Spektrum.de. Spektrum.de is the website of the popular-science magazine "Spektrum der Wissenschaft" (*spectrum of science*). Wikipedia is one of the largest curated open corpora on the Internet and is typically included in the training set of large language models. The category annotation of Wikipedia articles enables filtering content by domain. We decided to additionally use articles from Spektrum.de because they often use a more engaging and conversational tone, more akin to the spoken interaction that students are exposed to during lectures at university.

To choose relevant Wikipedia articles, we searched the German Wikipedia. We extracted level 2 technical terms from PoTeC stimuli and used them as search terms. From the resulting categories, we hand-picked 35 (e.g. botany, chronobiology, zoology[2]) for the biology corpus, and 32 (e.g. astrophysics, quantum mechanics, thermodynamics) for the physics one. The python library `wikipediaapi` was used to collect the names of subcategories of these categories. We used the `beautifulsoup` and `requests` libraries to scrape all articles belonging to the (sub)categories, yielding over 15 thousand articles for biology, and about 11 thousand articles for physics. The articles range

[1] The correlation coefficients between the three measures in the PoTeC dataset after removing words skipped in the first pass are: $r = 0.15$ ($p < 2.2e-16$) between first-pass and go-past time; $r = 0.59$ ($p < 2.2e-16$) between first-pass and total time; $r = 0.20$ ($p < 2.2e-16$) between go-past and total time.

[2] The search terms and categories were of course in German, but here the English equivalents are provided for easier reading.

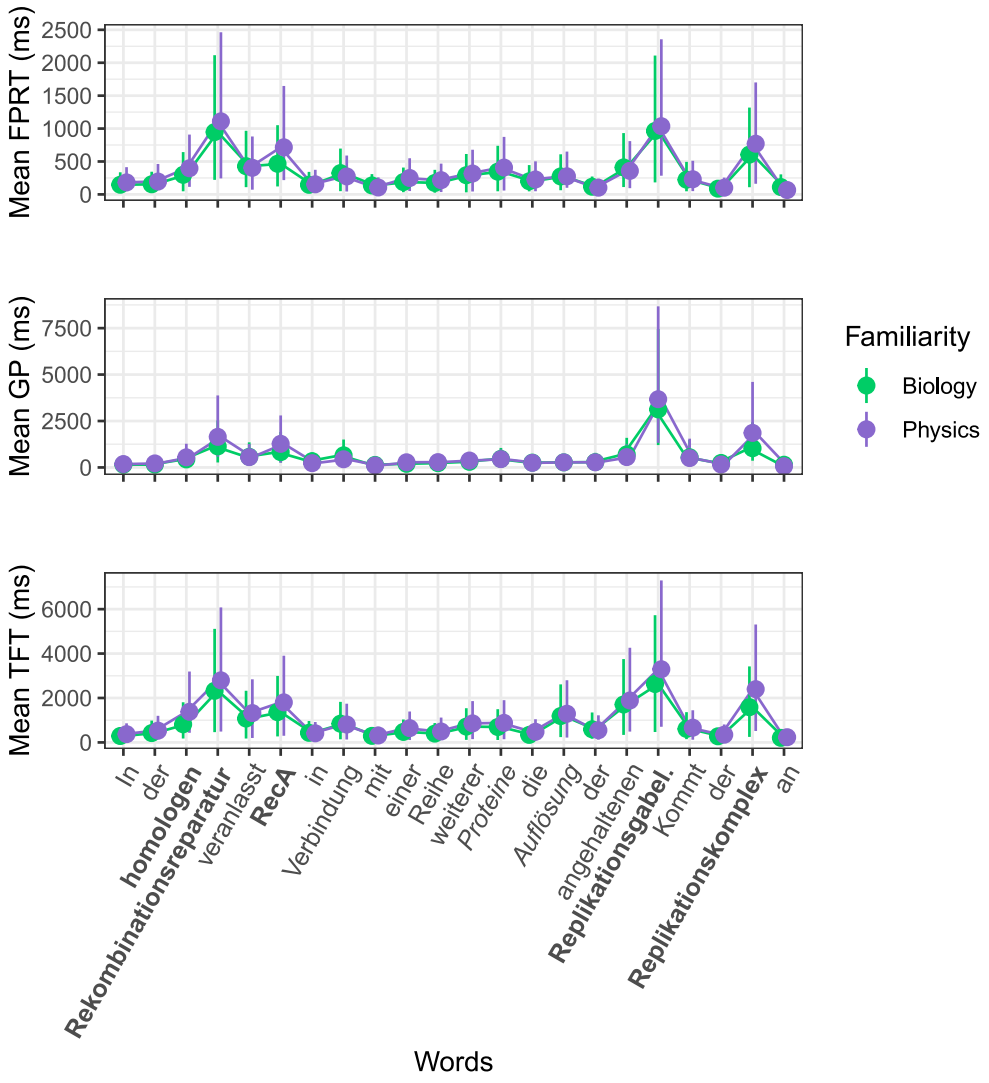

Fig. 1. Example sentence from the *b1* biology text with reading measures averaged across the two groups of readers given their domain familiarity. The error bars indicate the 95% confidence interval of the mean. The plot shows the averages for first-pass reading time (FPRT), go-past time (GP) and total fixation time (TFT) in milliseconds. The words in italics are technical terms of level 1, while boldface indicates a technical term of level 2. Common words are unmarked. The gloss with the translation of the text is provided in Appendix A.

Table 1
The number of documents, tokens and subwords (as yielded by the GerPT pretrained tokenizer) of the German domain-specific corpora for biology and physics from Wikipeda and Spektrum.de.

Domain	Corpus	Documents	Words	Subwords
Biology	Wikipedia	15,878	8,510,891	13,837,251
	Spektrum.de	9,103	2,084,463	2,893,905
Physics	Wikipedia	11,342	5,857,997	9,632,579
	Spektrum.de	2,820	609,202	852,978

from encyclopedic over scientific articles to biographies. Wikipedia articles needed additional cleaning and removal of math symbols.

The content on Spektrum.de is categorized into disciplines, among them also physics and biology. We scraped the news and columns of these categories, using the `requests` library to gain access, and `beautifulsoup` to parse the articles. In total, we collected about 9 thousand articles for biology, and almost 3 thousand for physics.

To create a corpus for each domain, we joined the respective documents from Wikipedia and Spektrum.de. Table 1 shows the final sizes of the domain-specific corpora. We created a fixed train (95%) and held-out test split (5%).

To assess their similarity, we compare the overlap in vocabulary between the collected corpora, PoTeC stimuli and a domain-general German corpus, the German portion of CC-100 (Conneau et al., 2020). CC-100 was used to train the German GPT-2 model used in our work. The corpus comprises Wikipedia and other online sources. We consider a sample of 120,000 documents from CC-100 for overlap calculation and use python libraries for that. The vocabularies of all corpora were lemmatized using `spaCy` (Honnibal et al., 2020) with stopwords removed using `nltk` (Bird & Loper, 2004). The list of stopwords includes determiners, prepositions, connectives, pronouns, question words as well as auxiliary and modal verbs.

Fig. 2 shows the overlap of lemmatized vocabularies across the five corpora. For example, the figure indicates that nearly 80% of the words from PoTeC biology texts are included into the sample of the general vocabulary in CC-100. In turn, only 0.02% of words from CC-100 are found in the PoTeC biology texts. We observe that the PoTeC texts are the most dissimilar to all other corpora, but they are closest to their domain-matching WikiSpektrum counterparts. We separately counted the number of overlapping technical terms between biology and physics PoTeC texts, finding 4 in total (*negativ* 'negative', *Energie* 'energy', *Ladung* 'charge', *elektrisch* 'electric').

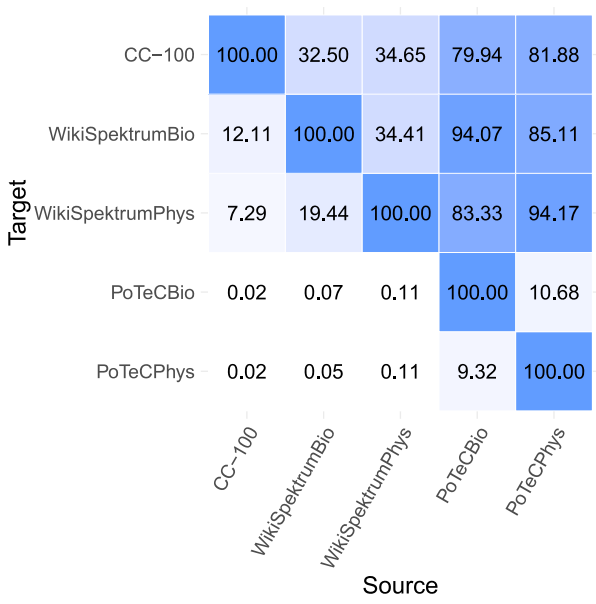

Fig. 2. Vocabulary overlap of lemmas in % between the PoTeC reading materials, the domain-adaptation train sets of Wikipedia and Spektrum, and a subset of the CC-100 train set. The lemmatized vocabulary of each corpus was created by removing stopwords, numbers and punctuation. Each cell indicates what proportion of the source corpus vocabulary is included in the target corpus vocabulary.

Study 1: Does experience in terms of domain familiarity affect reading?

This study explores the effect of domain familiarity on language comprehension, specifically on three measures of reading times. Previous studies of the PoTeC dataset have found that domain experts read faster than non-experts (Jakobi et al., 2025; Škrjanec et al., 2023), so this study serves as a replication of those results and a base for studies 2 and 3.

Methodology

We analyze the effect of domain expertise on three reading measures: first-pass reading time (FPRT), go-past time (GP), and total fixation time (TFT). For our analyses, the three measures are log-transformed and used as response variables in word-level linear mixed-effects regression models fit using the `lme4` package (Bates et al., 2015). The `lmerTest` package (Kuznetsova et al., 2017) was used to calculate p-values with the Satterthwaite method. We consider the following set of predictors:

- Word length as the number of characters without punctuation.
- Word position in text.
- Normalized word frequency of the lemma form estimated based on dlexDB.
- Terminology: a binary indicator as to whether a word is a technical term (common or technical).
- Expertise: a binary indicator about the reader's background (expert or novice).
- Interaction between terminology and expertise.
- Surprisal.

Word length, word position and lemma frequency serve as covariates. The frequency is taken from the dlexDB database (Heister et al., 2011) and we smoothed it by adding 1 to each value and then log-transformed it. The binary predictor *terminology* indicates whether a word is a technical term. The PoTeC materials distinguish between common words (level 0), and general (level 1) and expert (level 2) technical terms (Jakobi et al., 2025). Even though general technical terms are not specific to a single domain, they co-occur with expert technical terms, which means that domain knowledge is required for their comprehension in context. For this reason, we merge general and expert terminology into one group, resulting in two final levels (common words and technical term). The predictor *expertise* is a reader-specific predictor based on the combination of the reader's major (biology or physics) and the text domain (biology or physics). The predictor takes either the value "expert" (i.e. a physics student reading a physics text or a biology student reading biology) or "novice" (remaining combinations). The random effects in the regression models include a by-participant intercept, a by-word intercept and a by-word slope for the expertise level. Model versions with a richer random effect structure, for example including a slope for terminology, did not converge.

All continuous predictors were scaled and centered. The binary predictors are sum-coded: terminology (−1 "common", 1 "technical") and expertise (−1 "novice", 1 "expert"). For each of the three reading measures, we fit a baseline regression model (Eq. (2)).

$$\begin{aligned} LogRT \sim & Length + LogFreq + Position + \\ & Expertise * Terminology + \\ & (1|SubjectID) + (1 + Expertise|WordID) \end{aligned} \tag{2}$$

Results

The results are summarized in Fig. 3. There is a main effect of word length and position in the sentence: longer words and words that appear earlier in text have longer reading times in all three measures of reading duration. There is a significant main effect of word frequency, but only for go-past and total reading time.

Reader expertise also affects overall reading speed: on all three measures, reading times are lower for experts than for non-experts. We furthermore find a main effect of terminology in all measures: domain-specific terminology is read more slowly than common words. Additionally, we find a significant interaction between expertise and terminology: domain experts read technical terms faster than novices (i.e., they have less of a reading time slowdown on technical terms compared to novices). The full regression table can be found in the appendix (Table B.2).

To better understand the effect size of expertise, we transformed them from log to linear space in milliseconds. The effects and their 95% confidence intervals correspond to −11[−13, −8] ms for first-pass reading time; −23[−27, −18] ms for go-past time; and −81[−85, −77] ms for total reading time. The span of the confidence intervals indicates that the effects can be clearly detected.

Discussion

These observations are in line with previous findings on reading less familiar or unknown words (Chaffin et al., 2001; Lowell & Morris, 2014; Williams & Morris, 2004), which have reported slower reading and more regressions on these words. Previous work also found that the behavior of experts and novices is different in the early measure of first-pass RT, indicating that lexical access is impeded for novices: this may happen because the readers' vocabulary either does not include this word or because the word's meaning is not effectively pre-activated based on the previous context. Difficult words can additionally trigger regressions to preceding context for revision, confirmation or integration: go-past durations have been found to be longer for novices as well, meaning they are slower at moving on through the text. This leads to longer total reading time durations as well. Comparable effect sizes for the PoTeC have been reported in the study that introduced the dataset (see Figure 4 in Jakobi et al., 2025). Their analysis considered a more detailed account of expertise and split participants given the stage of

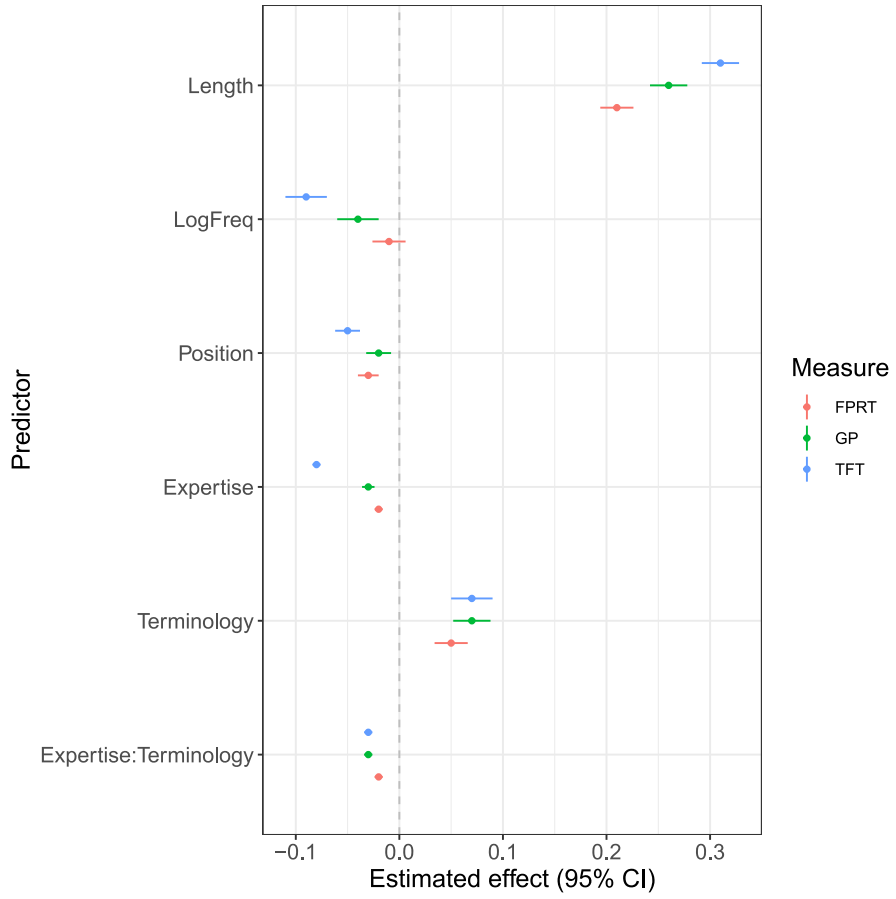

Fig. 3. The estimated effect sizes and their respective 95% confidence intervals on the three reading measures: first-pass reading time (FPRT), go-past time (GP) and total reading time (TFT).

studies (undergraduate and graduate) along with their domain (biology and physics), but, unlike our approach, their reading time model did not include the interaction between expertise and terminology. Despite these differences we observe similar patterns with respect to background knowledge.

Study 2: Predictive power of reader-aligned surprisal

The level of domain expertise affects the processing effort, surfacing as differences in reading times. In this study we address the question of modeling the processing effort of readers given their domain knowledge, including surprisal as a measure of processing effort in context. To estimate surprisal we use a pre-trained German Transformer language model, aligning it to the reader by adapting it to the domain of expertise of the reader (either biology or physics). This study explores what technique of adaptation and what amounts of training data result in more accurate estimations of processing effort in the evaluation against reading times. Two domain adaption techniques are used: one with updates to all weights of the language model with respect to the target domain (full fine-tuning), and one with leaving the pre-trained parameters intact and learning new, additional weights (adapters) that are trained to combine the existing knowledge according to the target domain.

In our analysis of each reading measure, we jointly fit one regression model for both groups of expertise, assigning biology students the surprisal estimated from a biology-tuned LM, while a physics model is used for physics students. In this, we explore all possible combinations of amount of training data and the domain, meaning some combinations include surprisal from the biology and physics models trained on similar amounts of data, but some combinations are more asymmetrical with one model fully domain-adapted, while the other is still at the start of training. This allows us to explore the different levels of domain adaptation and the readers' expertise without assuming that these are necessarily the same for adaptation to physics and biology domains.

Methodology

Language model choice and domain adaptation. To model the predictability effects during reading of German scientific texts, we use GerPT2[3] (Minixhofer, 2020), which was trained on German documents and is based on the GPT2[4] architecture. GerPT2 is an instance of GPT2-small with 163 million trainable parameters. To train GerPT2, the German section of the CC-100 corpus of web crawl data was used. While larger German language models are available, multiple studies (Oh & Schuler, 2023b; Shain et al., 2024) have shown that smaller Transformer models, of the size of GPT-2 in particular, yield surprisal estimates that predict reading times better than larger models.

We perform domain adaption of the GerPT-small model with two different methods: via full fine-tuning of all parameter weights, and by adapter training. To fine-tune the model fully, we use the `transformers` python library and continue training the model with the causal modeling objective.[5] We adapt the model to the biology and physics domains separately. To train bottleneck adapters, we use the

[3] https://huggingface.co/benjamin/gerpt2.
[4] https://huggingface.co/openai-community/gpt2.
[5] https://github.com/huggingface/transformers/tree/main/examples/pytorch/language-modeling.

`adapters` library[6] in python, setting the reduction factor hyperparameter to 16, meaning the size of the word embeddings (hidden size) is divided by 16 to obtain the adapter size of 48 (768/16 = 48). For both techniques, we use the following hyperparameter setting: a batch size of 8, learning rate of 1e−4, 100 warm-up steps, and training maximally for 16,384 steps without early stopping.

We are interested in the time-course of domain adaptation with respect to the fit to reading times (RT) and LM quality. To do that, we consider intermediate model checkpoints during training and save the model checkpoints every 4^n steps with $n \in \{1, 2, 3, 4, 5, 6, 7\}$, meaning the checkpoints lie at 4, 16, 64... 16,384 steps. In a single step, the model's or adapter's weights are updated based on 8 samples, which means we consider the models after seeing 32, 128, 512... 131,072 samples. We evaluate the domain-adapted models in terms of their linguistic accuracy (perplexity on held-out test datasets) and psychometric predictive power (using estimated surprisal for reading time fit).

Perplexity (see Eq. (3)) is defined as the inverse probability of the test data T given the language model, normalized by the number of words N in the test set. Lower values indicate that the test data is assigned a high probability by the language model, which is understood to reflect the model quality. If the perplexity value is high, the text data is less predictable or less likely according to the model's learned probabilities.

$$perplexity(T) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i|w_1 \ldots w_{i-1})}} \tag{3}$$

We test the language models on in-domain data (i.e. a model adapted to biology on the biology test set) as well as on out-of-domain texts (i.e. a biology model on physics data). In-domain testing evaluates the rate of adaptation to the domain, while out-of-domain perplexity indicates the model's ability to generalize outside of its specialized domain and to transfer the learned representations onto other domains.

Surprisal estimates. The GerPT2 model operates over subword tokens, which means words can be split into multiple subwords. To obtain word-level surprisal estimates, the PoTeC texts were first tokenized with the GerPT2 tokenizer. After subword-level surprisal was calculated, the surprisal scores of a word's subwords were summed to obtain word-level estimates, following the chain rule of probability. Due to the sensitivity of language model training to the initial state of weights, we train each model combination (domain adaptation technique × domain) with three different random seeds. Their per-word averaged surprisal estimates are used in the regression analysis. We consider the combinations between the two domains and the number of training steps when including surprisal into the model. For example, when fitting first-pass reading time with surprisal from fully fine-tuned language models, 49 averaged surprisal sources are considered (seven for biology and seven for physics).

We add surprisal as a fixed effect to the baseline regression model to create experimental models (Eq. (4)).

$$\begin{aligned} LogRT \sim & Length + LogFreq + Position + \\ & Expertise * Terminology + \\ & Surprisal + \\ & (1|SubjectID) + (1 + Expertise|WordID) \end{aligned} \tag{4}$$

Comparison of nested regression models. The goodness of fit of the model is evaluated via the likelihood ratio test by comparing the log-likelihoods of an experimental model (surprisal as one of fixed effects) and a baseline model (no surprisal). The difference is termed ΔLL and it is the result of subtracting the baseline model's log-likelihood from that of the experimental model (Eq. (5)).

$$\Delta LL = LL_{experimental} - LL_{baseline} \tag{5}$$

[6] https://github.com/adapter-hub/adapters.

The baseline and the experimental model differ in one predictor: surprisal is included in the experimental model as a main effect, but not in the baseline model. The likelihood ratio tells us whether it is beneficial to add surprisal to the baseline model. Adding surprisal from pre-trained neural language models typically improves over the baseline (Oh & Schuler, 2023a; Shain et al., 2024). In order to quantify the magnitude of this benefit, we calculate the difference in log-likelihood between the experimental and the baseline models. This allows for a comparison of different experimental models, each including surprisal estimates from a different language model (or, in our case, averaged surprisal across three LMs with different random seeds). The comparisons of the magnitudes of ΔLL are not to be interpreted as hypothesis testing, but rather an indication of numeric tendency: a positive ΔLL indicates the model fit is improved upon adding surprisal as a predictor; a larger ΔLL suggests a larger improvement over the baseline.

Results

Effects of domain adaptation on perplexity. Perplexity values with respect to the amount of training data are presented in Fig. 4. The plots show that domain adaption generally leads to lower perplexity. On in-domain data the perplexity drops consistently throughout training on the PoTeC text stimuli and the Wikipedia and Spektrum test sets across both domains. We take this as a sign of successful domain adaptation, i.e. the LM now more accurately represents the target domain.

Throughout domain adaptation, the perplexity on both out-of-domain test sets also decreases. Even when the LMs are tested in a cross-domain setting, it seems like learning on in-domain data benefits the out-of-domain representation. This is not surprising since both domains are branches of natural sciences and share some vocabulary (see Fig. 2). We also notice indications that out-of-domain generalization stops for the physics LM as its perplexity starts increasing again after 1024 steps for full tuning (on WikiSpektrum) and adapters (both test sets). For the fully fine-tuned physics LM on WikiSpektrum, the perplexity at the end of training again equals the perplexity at the very beginning of training.

A similar pattern is observed for the general Wikipedia texts. About half of the training data of the base language model for our study, GerPT2, was taken from Wikipedia (Conneau et al., 2020). While we specialized this LM to the biology and physics domains, the perplexity on the domain-general language reduced, but not below 30. After that, the perplexity started growing for all models (except for the biology adapter), which means that the LMs start to become so specialized to the target domain that the previously learned patterns of domain-general language are not predicted well anymore.

We observe interesting differences between the two adaptation methods in the trajectory of adaptation. At the beginning of training, fully fine-tuned models demonstrate a lower perplexity on all domains compared to adapter models. This can be explained based on how these methods work: in adapters, a randomly initialized new layers are introduced in between frozen pretrained layers. These randomly initialized parameters initially introduce noise which is not present in full fine-tuning, thereby leading to a comparatively higher perplexity. During the course of training, the new parameters are quickly trained, such that adapters also achieve lower perplexities in in-domain settings as training proceeds. Overall, the domain adaptation process requires more training steps in order to achieve similar levels of perplexity as full fine-tuning models. After the full set of 16 384 training steps, the in-domain perplexities of fully fine-tuned models are very low, at the cost of increased perplexities on out-of-domain data and domain-general data, indicating over-fitting of the fully fine-tuned models. At the same number of training steps, the adapter model on the other hand has achieved similar perplexity on all data sets as FFT models after 1024 steps — the adapter model is hence not trained to a point at which over-fitting effects become grave; first tendencies towards overfitting of the adapter on the general and out-of-domain data emerge for the physics model at the latest training steps.

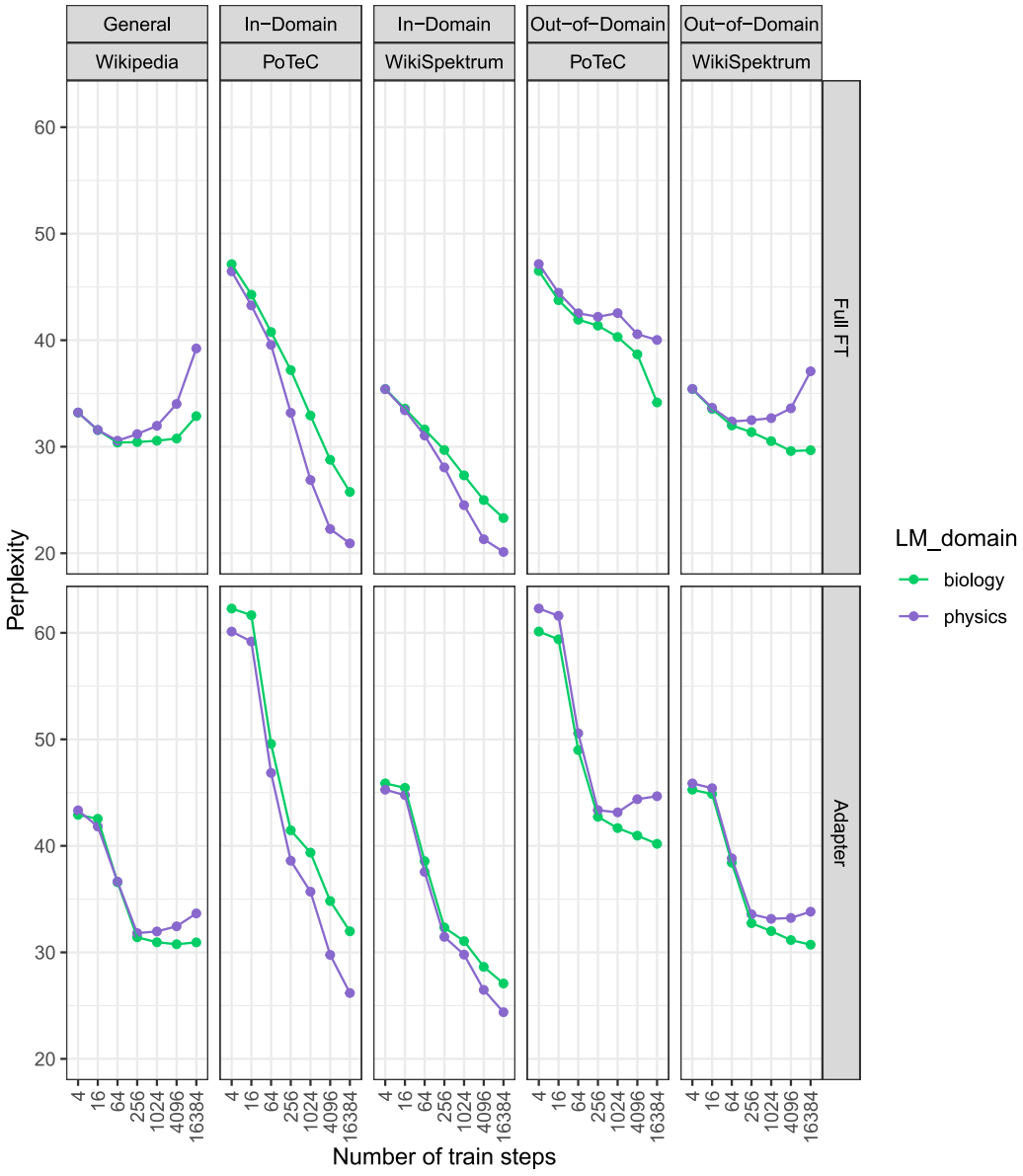

Fig. 4. Perplexity on test corpora. The perplexity values are averaged over 3 checkpoints of different seeds. The language models were adapted to the target domain either via full fine-tuning (*Full FT*) or with adapters (*Adapter*).

Effects of domain adaptation on surprisal. Fig. 5 illustrates how surprisal estimates for common and domain-specific terminology evolve throughout the process of domain adaptation via full fine-tuning and adapter training. The subplots also include surprisal estimates at step 0 of training, representing predictions made by the non-adapted base language model. Surprisal estimates for common words remain largely stable. In contrast, the surprisal of technical terms, especially for in-domain texts, decreases, indicating that the models are learning to assign these terms higher predictability. Some adaptation to technical terminology is also observed in out-of-domain data.

Similar to the perplexity trends shown in Fig. 4, surprisal estimates for the PoTeC texts also reflect that fully fine-tuned models adapt to the target domain more quickly: surprisal estimates for technical terms are lower in FFT compared to adapters for the same amount of training data. We can also see the effect of introducing the randomly initialized adapter layers on surprisal during early training: surprisal estimates are initially higher for adapter models with little training than for a general pre-trained model, and this effect is particularly pronounced for technical terms. After 64 steps of training, adapters improve their predictions of technical terms, and surprisal decreases beyond general-domain model surprisal estimates during further training. Fully fine-tuned models exhibit a steady, monotonic decrease in surprisal for technical terminology, in line with the overall perplexity results on in-domain and out-of-domain PoTeC data.

Predictive power of surprisal. The predictive power of surprisal for predicting first-pass (FPRT), go-past (GP) and total (TFT) time is depicted in Fig. 6, showing the results for language models that are either fully fine-tuned (*Full FT*) or adapters (*Adapters*). The color of the dots indicates a lower (yellow) or higher (orange to dark red) change in log-likelihood (ΔLL), i.e. predictive power. The number of fine-tuning steps for the biology and physics LMs is shown on the x- and *y*-axis, respectively. One dot represents the ΔLL of the reading time model with surprisal estimated from language models trained for the number of steps indicated on the x- and *y*-axis. For example, on the plot with fully fine-tuned models for first-pass time (FPRT; subplot A), the left-most dot in the top row shows the improvement over the baseline regression model upon including surprisal from a biology LM tuned for 4 steps (for biology students) and surprisal from a physics LM tuned for 16,384 steps (for physics students).

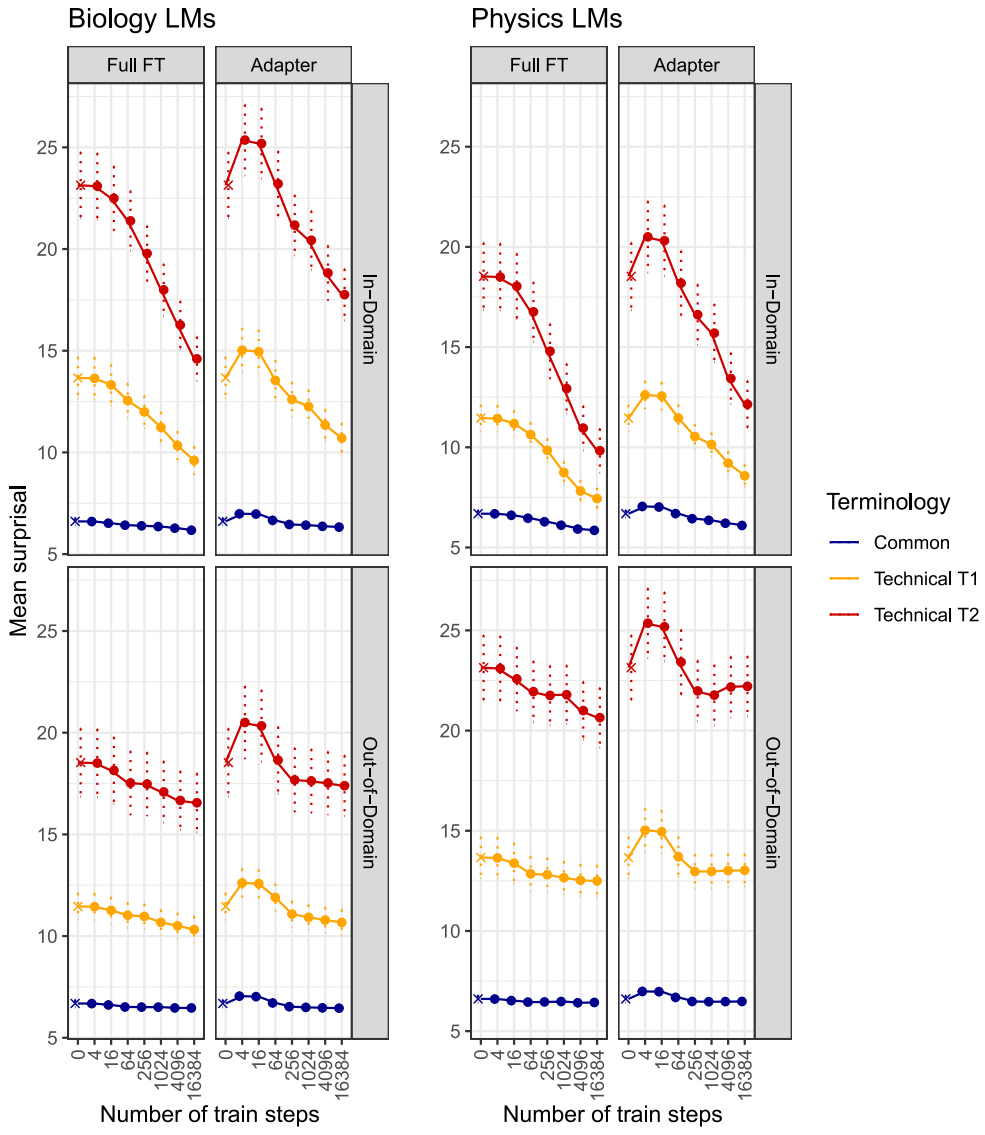

Fig. 5. Changes in surprisal of PoTeC texts throughout the course of domain adaptation, broken down by the terminology type and the match between language model and text domain (in- and out-of-domain). The error bars indicate the standard error of the mean.

Number of training steps. Across all LMs, we can observe that the improvement of model fit changes with the number of training steps, i.e. the amount of training data during adaptation to the target domain. We generally first observe an increase in model fit as training steps increase, which is followed by a drop in predictive power for very high numbers of training steps. These results hold for both the adaptation to the biology and the physics domains. However, the physics domain adaptation tends to need fewer steps than the biology domain.

Domain adaptation method. Subplot A presents the results for fully fine-tuned LMs, subplot B those for adapters. We had already seen in the perplexity and surprisal results that the parameter-efficient adapter method – in which fewer parameters are changed in each step – requires more training steps to adapt to the target domain. Here, we see that it also achieves optimal predictive power later than the model using full fine-tuning. In particular, we note that the surprisal estimates obtained via the adapter fine-tuning method on the biology domain might even benefit from more than 16 thousand training steps for predicting the reading times at late measures, as indicated by red dots at 16,384 training steps.

Effect on RT measures. The results exhibit a discrepancy between early and late reading measures: for first-pass time, the highest ΔLL is reached after 256 and 1024 steps (full fine-tuning and adapters, respectively). The optimal point of LM domain adaptation for predicting go-past and total reading time comes later: at about 1024 and 4096 (full fine-tuning and adapters, respectively). For the two late measures, the surprisal from fully fine-tuned models shows a decline in predictive power, indicating that this model may not only be overfitting with respect to out-of-domain data but also with respect to the surprisal experienced by humans. On the other hand, the surprisal estimated from adapters could potentially contribute to the fit even more if the training had been continued. We comment on the differences between reading measures in the discussion below.

The largest predictive power. Having outlined the general patterns of fit with reader-aligned surprisal, we now turn our attention to the point of maximum improvements for each reading measure. Numerically, at these points, surprisal contributes most to the quality of the fit. For the best fit of the early first-pass RT, surprisal is estimated with LMs at 256 training steps of adaptation (for both, biology and

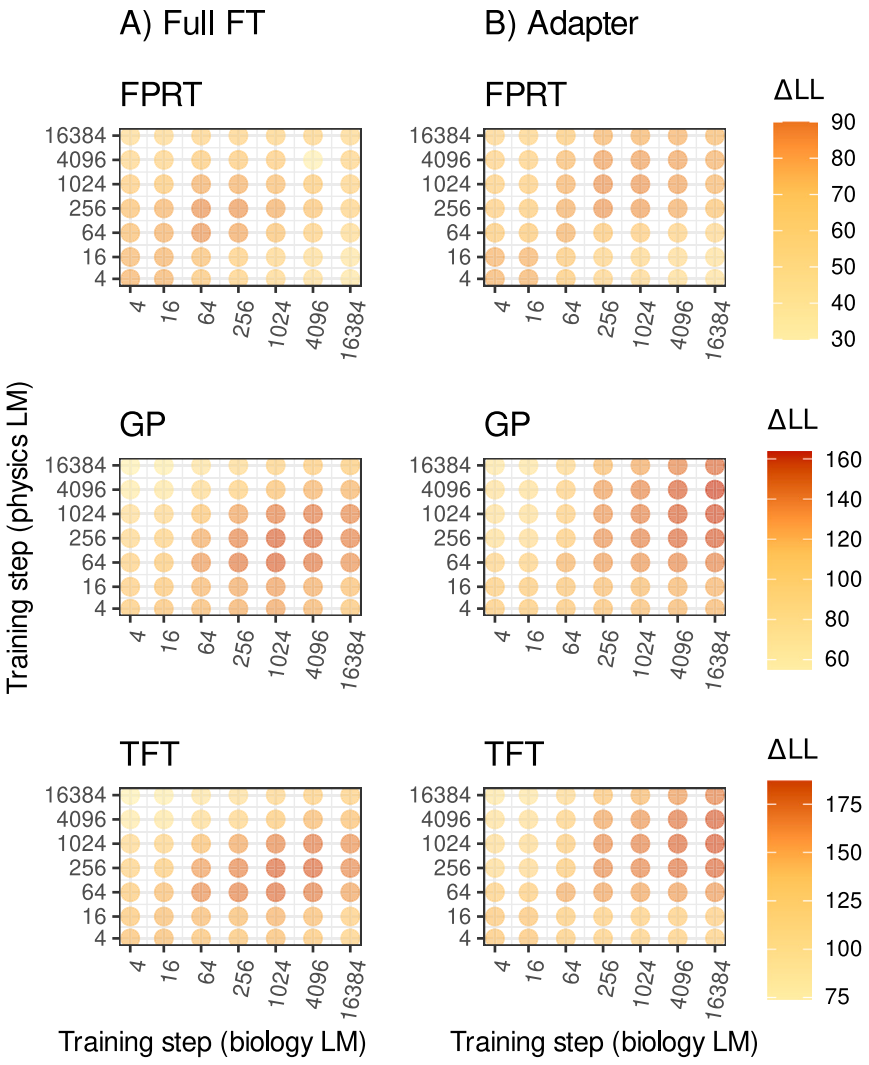

Fig. 6. Change in the log-likelihood (ΔLL) upon adding reader-aligned surprisal to the reading time regression model with surprisal estimates coming from either a fully fine-tuned model (*Full FT*, subplot A) or an adapter (*Adapter*, subplot B). The number of training steps of domain adaptation is indicated on the *x*-axis for biology LMs, and the *y*-axis for physics LMs. The plot shows results for first-pass (FPRT), go-past (GP) and total fixation time (TFT). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

physics) with a 85.8 improvement over the baseline. For both late measures, adapters also did best, after 4096 steps for both domains. The ΔLL for go-past time is 154.6, and 175.8 for total RT. The improvements are statistically significant (see Appendix C for details on significance testing). Our analyses in Appendix C furthermore demonstrate that a model including reader-aligned surprisal fits the data significantly better than a model that includes only general surprisal estimates from a GPT2 model that has not been domain-adapted.

Discussion

To understand the findings on the predictive power of surprisal, it is best to consider them in parallel with LM perplexity (Fig. 4) and average surprisal estimates (Fig. 5). We found that model perplexity improves on both in-domain and out-of-domain data for the first 256 steps of fine-tuning on either domain. Our reading time models also show that the first 256 iterations lead to an improvement in fit to reading times for all models. We can also see that continued fine-tuning on biology data leads to a continued improvement in perplexity on biology data, and a small improvement on physics data. Our results on surprisal estimates on the PoTeC data similarly showed continuously decreasing surprisal estimates for technical terms (and stable surprisal estimates for common terms) throughout fine-tuning. However, we do observe that these increasingly low perplexities and surprisal estimates do not entail a better fit to reading times: for the model fine-tuned on biology texts, we instead observe that model fit is best for 1024 and 4096 training steps, and explanatory power starts decreasing afterwards. This effect is even more pronounced for models fine-tuned on physics, where best predictive power is observed for FFT models trained for as little as 256 steps. Considering this observation together with the finding that perplexity on domain-general data also starts to increase at this level of fine-tuning, indicates that models are starting to overfit to physics data and exhibit catastrophic forgetting effects in the general domain. These observations are in line with the findings on the relationship of perplexity and predictive power of surprisal for domain-general models as reported in Kuribayashi et al. (2021) and Oh and Schuler (2023a). They report that as language model quality (as measured by perplexity) rises, the ability of its surprisal estimates to predict human reading times also increases, but only until reaching a point where the LM quality and the fit to human RT start to diverge —

the LM linguistic accuracy continues to improve with training, but this has an adverse effect on the predictive power of surprisal.

We observe that the two domain adaptation methods are comparable in terms of the predictive power of their surprisal estimates, that is, both models lead to fits of similar quality. Adapters offer a more parameter-efficient approach that needs fewer compute resources, requiring, however, more fine-tuning iterations to achieve a similar fit, cf. He et al. (2021), but also see Bansal et al. (2022).

In the full fine-tuning approach, the vector representations (embeddings) of words, the attention layers, the feed-forward networks, and the final layer are updated based on the error signal during pre-training on domain-specific data. In contrast, with adapters, the representations and relationships learned by the model during training on domain-general data are preserved and remain unchanged during domain adaptation. The purpose of training an adapter is to learn how to combine and weigh this existing knowledge to better align with the target domain. Previous research (van Schijndel & Linzen, 2018; Yan & Jaeger, 2020) has demonstrated that humans can easily and rapidly adapt their expectations to different domains. This behavior can in principle be modeled in different ways: while relatively small models like GPT-2 may not be able to represent the different contexts effectively and store different expectations, it is conceivable that larger models might be able to more effectively condition on different situational contexts and also exhibit highly adaptive behavior. Conceptually, the idea of adapters which can be used to model how a reader behaves in different situations is however also a practical modeling solution. We note however that both the implementation of inserting additional adapter layers and the behavior (worse prediction capabilities in early stages of training) lack cognitive plausibility. In future work, it would be interesting to measure whether the effect of extensive exposure to domain-specific texts influences human reading behavior when processing domain-general text, in line with the predictions of a monolithic fully fine-tuned model. However, our current data cannot speak to this question. Despite the conceptual and implementation differences between these two techniques, we overall observe more similarities than differences in the patterns of the predictive power of their surprisal across reading measures on the PoTeC dataset.

When humans acquire new knowledge, they also learn new words. How does this translate to LMs and the two adaptation techniques? It does not, at least not in a straightforward way. While LMs can approximate meanings for unknown words in ways that partially align with human inferences (de Varda et al., 2024), it remains less clear how novel words are learned and how this relates to their segmentation into tokens. The models used in this study rely on tokenizers based on Byte-Pair Encoding (BPE; Sennrich et al., 2016), a subword segmentation algorithm that uses unigram word frequencies to define merge rules. This allows LMs to flexibly represent out-of-vocabulary words by splitting them into known units. However, tokenization also introduces rigidity: the segmentation rules are fixed after tokenizer training and do not adapt during subsequent fine-tuning of the LM. As a result, even if a previously less frequent word becomes more frequent in a new domain, its tokenization remains unchanged. The tokenization is thus is a reflection of the pre-training data (Hayase et al., 2024). This rigidity has implications for estimating word probabilities and therefore surprisal. When a word is split into multiple subwords its probability estimate is smaller than if the same word would have been tokenized as a single subword (Lesci et al., 2025). This thus raises the question of whether a tokenizer should be adaptable.

Studies directly examining tokenization using behavioral measures of language processing are sparse. Some evidence suggests that a higher number of subwords correlates with slower reaction times in lexical decision tasks (Beinborn & Pinter, 2023), though in general, the choice of tokenizer (whether BPE, morphological, or whitespace) does not appear to significantly affect the predictive power of surprisal for reading times (Nair & Resnik, 2023).

During full fine-tuning, the representations and relationships of words across all layers are modified, tuning exclusively to the target domain, thereby overwriting previous representations — these only serve as a starting point for learning about the new domain. In contrast, adapters do not alter the word representations themselves but instead combine the existing representations with new adjustments. This process does not directly correspond to human language learning: when humans use existing knowledge to acquire new words and concepts (Gaskell & Ellis, 2009), their word-level lexicon does not remain static.

Another point of consideration raised by our findings is the observation that reading times in the early measure are best predicted by models with relatively small amounts of domain adaptation, while later reading time measures are better predicted by models with more intensive domain adaptation. We propose two possible speculative explanations for this finding. The first explanation relates to differences between expert and novice readers.

Study 1 shows that the effect sizes of the main effects of expertise, terminology and the interaction between them are smaller for the first-pass reading time than for the later measures. This means that the differences between expert and non-expert readers are not yet very large in the early reading measures. In later reading measures, which, respectively, count additional regressions from the word and later re-fixations of a word, novices show a more strongly different behavior from experts; in particular, they do regress and re-read more often. This behavior is captured by our reader-adaptive surprisal measure of fully-adapted domain-specific models: A reader will experience higher surprisal and regress more when reading text outside their domain of expertise, and experience lower surprisal and regress less inside their domain of expertise. Models that are domain-adapted more strongly can predict these differences more precisely than models that have been domain-adapted only for a very short period, and therefore are still quite similar to domain-general models which predict the equivalent reading times independent of a reader's personal expertise level. This idea is also consistent with the finding that reader-adapted surprisal estimates have significantly higher predictive performance than text-adapted surprisal estimates on later reading measures, while there is no significant difference between reader-adapted and text-adapted surprisal in first pass reading times (see Appendix C).

The second speculative explanation focuses on domain adaptation and the process of learning to predict domain-specific words. When LMs are trained from scratch to predict the next word given a context, they initially learn to predict the most frequent words, first capturing unigram frequencies and only later incorporating longer contexts (Chang & Bergen, 2022). We hypothesize that during further training on a specific domain, LMs prioritize learning the distributional properties of words that are most distinct from general language, so those that stand out in the domain. This suggests that domain adaptation emphasizes domain-specific words in the early stages of training, while semantic relations are acquired in subsequent stages, followed by commonsense knowledge and natural language understanding, which require the most exposure to training data (Zhang et al., 2021). We also know that fine-tuning primarily affects the upper layers of the model (Mosbach, 2023). Even in adapters, it is the final layers that steer the prediction towards the target domain (Alabi et al., 2024). As final layers of the LM are believed to handle word frequency (Kobayashi et al., 2023), these points taken together indicate that models might already capture frequency effects fairly well after only a few steps of domain adaptation.

Previous work also showed that frequency effects are typically present both in early and late reading measures, and that surprisal effects tend to have larger effect sizes in later reading measures (de Varda & Marelli, 2023; Greenberg, 2023; Shain et al., 2024). The idea that strongly domain-adapted models are less predictive of early reading times, possibly because their surprisal estimates primarily reflect frequency effects, while models trained for longer periods produce

more contextualized surprisal estimates that better predict later reading measures, appears consistent with previous findings.[7]

Study 3: What alignment of the language model is most beneficial for reading time prediction?

The surprisal estimates of domain-adapted language models can be included as predictors of reading times in different ways. On the one hand, the language model can be chosen such that it represents the training data best, i.e. given the text domain such that the physics LM is used to estimate surprisal of physics texts, and the biology LM for biology texts, regardless of the readers' expertise. We refer to this as text-aligned surprisal.

On the other hand, surprisal can be reader-aligned, with estimates coming from a language model that is matched with the readers' experience. While alignment with the text means that the language model has a low perplexity on this text (e.g. a biology-tuned LM on biology text), we hypothesize that the evaluation against reading times will show that reader alignment is a better method of choosing the language model because it better represents the processing effort given the domain experience.

Methodology

While the language models in Study 2 match the readers' domain expertise, this study analyzes the predictive power of surprisal from text-aligned language models. For biology reading materials the biology-adapted models were used to estimate the surprisal of both, biology and physics students. The surprisal for physics texts was calculated with physics LMs. Like in Study 2, we consider surprisal from LMs throughout the process of domain adaptation with either adapters or full fine-tuning. We evaluate the predictive power of text-aligned surprisal with ΔLL and visualize its values.

Results

Fig. 7 shows the predictive power of text-aligned surprisal from both domain adaptation techniques (full fine-tuning and adapters) throughout the process of training, for each of the three reading measures: first-pass (FPRT), go-past (GP) and total reading time (TFT). We generally notice that adding text-aligned surprisal has a beneficial effect on reading time prediction with the quality of the fit improving across reading measures and adaptation techniques. This is reflected in positive ΔLL values.

The quality of the fit seems to vary very little throughout the process of domain adaptation for biology and physics models: especially for go-past and total time, surprisal estimates based on various amounts of training data lead to very similar values of ΔLL. This contrasts with the results found for reader-aligned surprisal in Fig. 6. Both figures use the same legend settings for each reading measure making the figures directly comparable. In the reader-aligned results we see clear benefits of the alignment with the model fit achieving a higher quality than for text-aligned. This holds for the two late measures (GP and TFT), where surprisal estimates from models with both adaptation techniques with longer training have a larger predictive power than those from models with less training. This, in turn, does not seem to be the case for first-pass time: the results with text- and reader-aligned surprisal for fitting first-pass RT are almost indistinguishable, both showing a benefit to the same degree.

[7] We also point the interested reader to the discussion in Shain (2024) regarding the relationship of frequency and surprisal: their analyses indicate that frequency and surprisal effects are dissociable, i.e., that they cannot be reduced to one another in human processing, but also that effects of both of these metrics can be found in both early and late reading measures.

We conducted additional statistical tests to compare the fit quality of text- and reader-aligned surprisal focusing on the point of training with the highest ΔLL for each reading measure. The methodology and results are described in Appendix C. To summarize, reader-alignment indeed leads to a better fit than text-alignment for the two late measures. However, for first-pass time, the two alignment methods are comparable, with neither showing a clear advantage.

Discussion

The results for text-aligned surprisal show that the values of ΔLL stay in similar ranges around the median for each reading measure, both at the beginning and end of training. After 4 steps of domain adaptation, the LMs are hardly tuned to the domain: the perplexity of the full fine-tuning on in-domain PoTeC data is still reduced after subsequent further training steps, even more so for adapters (Fig. 4). After 1024 steps, the perplexity values are much lower indicating that the LM can indeed predict the words of the domain well. This is also reflected in surprisal values for technical terminology (Fig. 5): after 4 steps of training they are still high and close to those of the non-adapted LM (step 0), but drop throughout the course of domain adaptation. The level of domain adaptation seems to matter very little when all readers (regardless of their domain expertise) are assigned the same surprisal estimates, and the source of surprisal is aligned to represent the text best (i.e. is text-aligned). The amount of training neither improves nor harms the fit to reading times. In the course of training, LMs are tuned to the target domain and the properties of its language. For roughly half of the participants in the dataset, the reduced surprisal provides a more accurate estimation of their processing effort, while this is less true for the other half. We speculate that this balance between improvement for some participants and disadvantage for others results is a trade-off, leading to minimal variation in fit quality across different stages of LM adaptation.

The predictive power of text-aligned surprisal in Fig. 7 does not fluctuate much with training, but it lies consistently above the baseline and significantly improves the fit (see Appendix C for significance testing). This indicates that estimating the same surprisal model for all readers can still be informative for reading time prediction. In studies where the domain familiarity of the readers was not assessed or is otherwise unknown, it is still a good idea to estimate a domain-adapted surprisal model. However, we find that an even better modeling choice is to use reader-aligned surprisal: when the domain knowledge was considered in estimating the processing cost in terms of surprisal, the models of reading time were of higher quality.

General discussion

Modeling of other cognitive constructs

Our study investigates the effect of domain expertise on eye-movements during reading as a group-level difference between readers, and observes that expertise has effects that can be found both in early and late reading measures. Expertise is however not the only dimension along which readers differ — other cognitive properties or abilities might affect reading, due to effects on the retrieval of relevant concepts, or text comprehension and recall. For example, working memory capacity (Kaakinen et al., 2003; Ricks & Wiley, 2009; Schurer et al., 2020) has previously been shown to have effects on reading. Computational models that mimic working memory capacity effects in humans could hence be used in order to model human readers with different working memory capacities, and test whether such memory-capacity matched models can predict reading times more accurately, compared to generic models. In fact, there are recent first attempts which indicate that memory-limited large LMs have better the psycholinguistic predictive power for explaining reading times compared to current models that

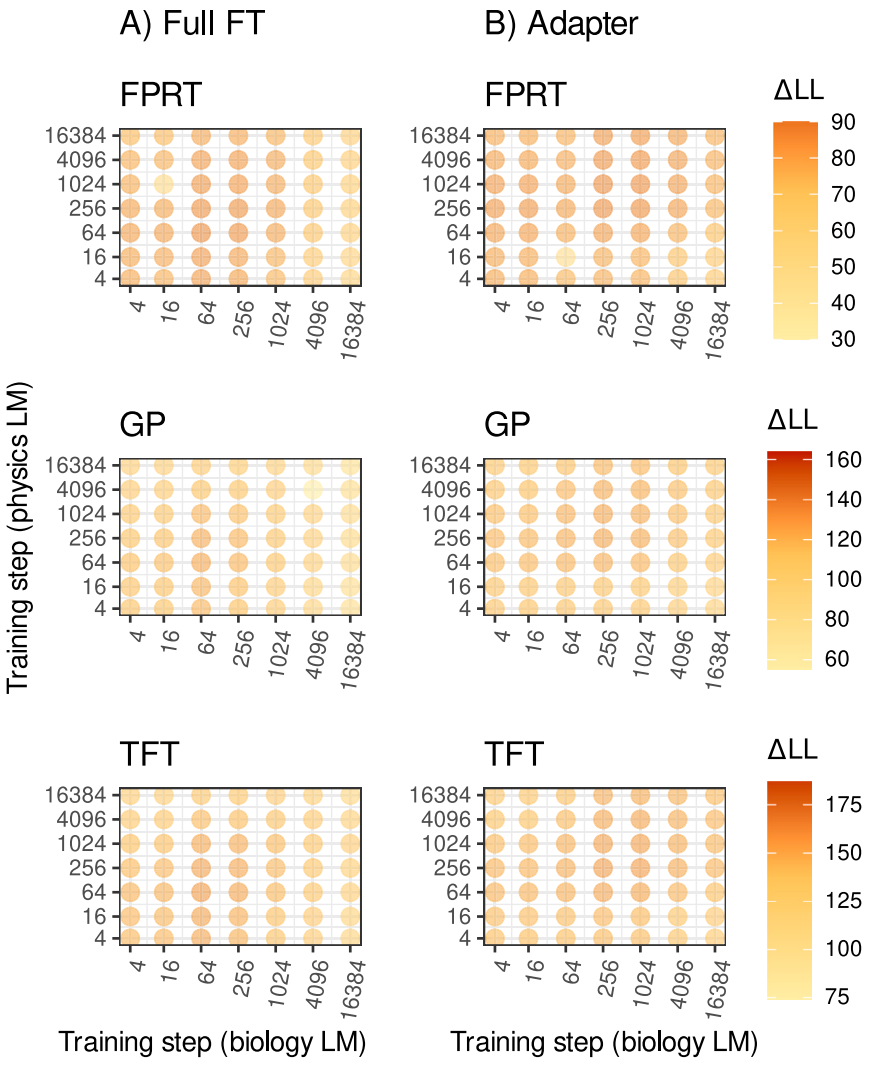

Fig. 7. Change in the log-likelihood (ΔLL) upon adding text-aligned surprisal to the reading time regression model with surprisal estimates coming from either a fully fine-tuned model (*Full FT*, subplot A) or an adapter (*Adapter*, subplot B). The number of training steps of domain adaptation is indicated on the *x*-axis for biology LMs, and the *y*-axis for physics LMs. The plot shows results for first-pass (FPRT), go-past (GP) and total fixation time (TFT).

have unrealistically high memory capacity (Kuribayashi et al., 2022; Timkey & Linzen, 2023).

However, these works have only been modeling population-level effects, and could be extended to individual-level or group-level predictions in future work. To test hypotheses about the effect of cognitive factors and whether they can be successfully modeled in computational models, it is necessary to validate predictions against reading corpora that also have information on readers' individual differences such as their working memory capacity or literacy level. The recently developed InDiCo corpus (Haller et al., 2024-11-28) of eye-tracking-during-reading includes several tests of individual differences, and could be a useful resource for future research in this direction.

Informativeness and meaning of the eye movement signal

We find that domain expertise has a main effect on early (first-pass reading) as well as late measures (go-past and total reading time), but find that the alignment of surprisal patterns with the measures differently. Most notably, the prediction of first-pass RT benefits from the reader- and text-aligned surprisal predictor equally well, while reader-aligned surprisal shows to be more informative than text-aligned for both late measures. We speculate that this is either related to the domain specificity of the surprisal estimates or to the possibly distinct cognitive processes that the reading measures index. The measures are related and even included in each other (i.e. first-pass RT is included in go-past as well as total RT), but we assume they index different processes since they lead to different results. A promising first step is to explore the sensitivity of each measure to oculomotor and lexical processing-based factors (Heilbron et al., 2023).

One empirical approach to disentangle the functions of different reading measures is using co-registration to simultaneously record the eye movement signal and, for instance, event-related potentials (Frank & Aumeistere, 2024; Hollenstein et al., 2018). This can provide a key for the better understanding of the eye-mind link and the temporal dynamics of language processing during reading. For example, van Moort et al. (2020) found diverging results from the eye movement and the fMRI signals in the study of reading text that contradicts prior knowledge: while their behavioral data from eye-tracking shows inconsistent patterns for text- and knowledge-based contradictions, brain imaging suggests that text-based and knowledge-based information during reading are processed differently. Our study uses data from

naturalistic reading. However, an experimental design with clearly defined conditions and precise predictions about the measures could complement our findings and provide more clarity on the specific cognitive processes that they reflect.

In line with this research agenda, several studies have developed theoretical accounts of specific reading measures (Andrews & Veldre, 2020; Paape & Vasishth, 2022; Weiss et al., 2018; Wilcox et al., 2024). The extent to which word predictability influences these measures and the mechanisms by which predictability should be estimated has been explored by de Varda et al. (2023). Their findings indicate that both human-based and computational estimates of predictability can successfully predict various reading time measures, albeit with differing levels of accuracy, leaving room for further theory building.

Computational models between theory and technology

In our study, we find that reader-alignment of LMs results in better estimations of human processing effort than text-alignment. However, even text-aligned domain-adapted models are shown to outperform domain-general surprisal models in terms of their ability to predict reading times. For the PoTeC corpus, the improvement in reading time prediction from text-aligned surprisal may be due to the fact that it still aligns with the reader's experience for half of the data, and because the two domains used in the corpus share overlapping vocabulary. Our results do not support the conclusion that text-aligned surprisal would also predict reading times better than a domain-general model if the readers were all novices, so that the domain-general model would in fact model all readers adequately. In fact, this would be an interesting and testable prediction for future work.

The domain adaptation methods that we have proposed here have proven to be computationally adequate models in that they capture the differences in reading patterns between domain experts and novices. However, these findings at the computational level do not warrant conclusions about the architectural level. In fact, the adapter models, which ultimately exhibited numerically the best results with respect to predictive power of the reading times are cognitively implausible in many ways: they involve introducing additional randomly initiated variables into the network, and as a result, during the first steps of domain adaptation, their prediction accuracy on general text as well as the new text decreases. This is implausible both from an architectural point of view and from a performance point of view. While full fine-tuning shows more plausible behavior, it is computationally expensive and exhibits catastrophic forgetting effects on the general domain. Future work could investigate such effects in more detail to better understand how readers adapt their expectations to different contexts or situations.

More generally, this raises a question on the role of language models and their development: language models as language tools vs. implementations of theories supporting computational cognitive modeling. Guest and Martin (2021) caution against a prevalent fallacy in cognitive science: inferring that because a computational model correlates with or predicts human data, it must operate via the same mechanisms as human cognition. This inference is logically flawed and Guest and Martin (2021) point out two primary concerns: one regarding models, the other concerning data.

First, functional correspondence does not entail mechanistic equivalence. Distinct underlying mechanisms can yield similar input–output behavior, which is a principle known as multiple realizability (e.g. Fodor & Pylyshyn, 1988). This is especially pertinent in frameworks like surprisal theory, which provide predictive success without specifying generative mechanisms. As demonstrated by comparable results across different domain adaptation techniques, similar predictive performance does not imply shared cognitive processes. Neglecting this distinction risks misattributing cognitive validity to models and mischaracterizing human cognition itself.

Second, while empirical data is essential for theory evaluation, it is not sufficient for theory construction. As Guest and Martin (2021) argue, data collection is shaped by theoretical assumptions, and thus empirical results must be interpreted within a rigorous theoretical context. The rapid deployment of large language models enables quick correlation analyses with human data, but often lacks the theoretical scrutiny necessary for meaningful cognitive insight.

Additionally, van Rooij et al. (2024) argue that AI's current framing as a technological endeavor may detract from its value in advancing theoretical understanding of cognition. Nonetheless, the field holds promise. Advancing computational cognitive modeling requires developing interpretable models informed by cognitive theory (van Rooij et al., 2024), and characterizing the internal operations of deep neural networks in psychologically meaningful terms (Guest & Martin, 2021). While both human cognition and neural networks remain partially opaque, the recent development of interpretability methods offers potential for using such models as valuable tools for cognitive theory (McGrath et al., 2024).

Conclusion

In this study we focus on adapting language models to better account for differences between readers by improving the predictive power of surprisal estimates. Specifically, we analyze the early and late measures of reading times of participants with a background in either physics or biology. We add to the body of work on the use of language models for modeling online language comprehension by exploring two methods for domain adaptation: full fine-tuning and adapters. As opposed to general language models or those adapted to the domain of the text, we find that aligning the models to the expertise of the reader led to better surprisal estimates for predicting late measures, while the early measure shows less clear results. Late measures are an aggregate and as such might be more robust to unaccounted individual differences, e.g. in reading styles. Our results leave open the question what processing cost surprisal operationalizes in different reading measures over the time course of reading. Comparing fully fine-tuned and adapter models, we find similar empirical results in predictive power of surprisal. Adapters are more parameter efficient than full fine-tuning, making them a promising tool for computational modeling. However, to support theory building, it is crucial that future work looks beyond the model behavior and considers the mechanisms that drive this behavior as well as revises engineering decisions such as tokenization that affect our conclusions about word representation in language models.

CRediT authorship contribution statement

Iza Škrjanec: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Vera Demberg:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank Mikhail Sonkin for the help with developing language models. They also thank Marian Marchal and Merel C. J. Scholman for the valuable discussion about the initial results. The authors also thank the reviewers and editors for their constructive feedback that helped to improve the manuscript. The work reported in this paper has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Germany – Project-ID 232722074 – SFB 1102. Iza Škrjanec is supported by the Konrad Zuse School of Excellence in Learning and Intelligent Systems (ELIZA) through the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research, Germany.

Appendix A. Example stimulus

Translation of the example stimulus from Fig. 1.

(1) In der homologen Rekombinationsreparatur veranlasst RecA in Verbindung mit einer Reihe weiterer Proteine die Auflösung der angehaltenen Replikationsgabel. Kommt der Replikationskomplex an
In the homologous recombination-repair triggers RecA in conjunction with a range other proteins the dissolution the arrested replication-fork. Arrives the replication-complex at
'In homologous recombination repair, RecA, in conjunction with a number of other proteins, triggers the dissolution of the arrested replication fork. When the replication complex reaches'

Appendix B. Linear mixed-effects model from Study 1

See Table B.2.

Table B.2
Regression coefficients and test statistics from the baseline linear mixed-effects models for three reading measures: first-pass (FPRT), go-past (GP) and total fixation time (TFT). The asterisk indicates the p-value: *** ($p < 0.001$), ** ($p < 0.01$), * ($p < 0.05$), no asterisk ($p > 0.05$).

Measure	Variable	β	SE	t	p
FPRT	(Intercept)	5.62	0.02	334.65	***
	Length	0.21	0.008	28.02	***
	LogFreq	−0.01	0.008	−1.42	
	Position	−0.03	0.005	−5.34	***
	Expertise	−0.02	0.002	−11.66	***
	Terminology	0.05	0.008	5.99	***
	Expertise:Terminology	−0.02	0.002	−8.18	***
GP	(Intercept)	5.95	0.02	266.91	***
	Length	0.26	0.009	28.07	***
	LogFreq	−0.04	0.01	−4.01	***
	Position	−0.02	0.006	−3.02	**
	Expertise	−0.03	0.003	−13.31	***
	Terminology	0.07	0.009	7.69	***
	Expertise:Terminology	−0.03	0.002	−10.25	***
TFT	(Intercept)	6.23	0.03	191.90	***
	Length	0.31	0.009	33.65	***
	LogFreq	−0.09	0.01	−8.33	***
	Position	−0.05	0.006	−8.97	***
	Expertise	−0.08	0.002	−34.41	***
	Terminology	0.07	0.01	6.82	***
	Expertise:Terminology	−0.03	0.002	−11.80	***

Appendix C. Comparison of alignment methods

It is possible to think of language models as general tools that can be adapted to represent either a domain or a group of readers better. To synthesize the observations from Study 2 and this study, we use a statistical test to compare the quality of the predictions of reading times with models including either 1) general surprisal from a domain-general language model (the original GerPT2), 2) text-aligned surprisal from a domain-adapted LM, 3) reader-aligned surprisal from the same domain-adapted LM.

We use the domain-general GerPT2 as is. For surprisal from domain-adapted models, we use language models that had the highest improvement over the baseline in terms of ΔLL for each reading measure: an adapter LM at 256 steps for first-pass-reading, and adapters at 4,096 steps for both go-past and total reading time. This selection procedure allows for a small number of comparisons, i.e. avoiding the problem of multiple comparisons. We also emphasize the test results are to be interpreted in conjunction with the trends visualized in Figs. 6 and 7.

To compare the fit of non-nested models, we apply the Vuong likelihood ratio test (Vuong, 1989) for pair-wise comparisons using the `nonnest2` package in R (see also Merkle et al., 2016). The test compares the likelihoods of the data under the two models, correcting for any differences in the degrees of freedom.

Due to technical issues with memory, we were not able to compare two linear mixed-effects regression models with the Vuong test. To still compare the quality of the fit given surprisal predictors, we fit simple linear regression models, with the same fixed effects as the linear-mixed effects counterparts, but without the random effects structure. While this reduces the complexity of the modeled data, we resort to linear models only for comparison of non-nested regression models. b provides the effect sizes, standard errors and test statistics for mixed-effects models and their basic regression counterparts, showing that the simple regression and the linear mixed-effects model estimates are very similar.

Table C.3 shows the raw log-likelihood alongside with ΔLL (with respect to the baseline), the χ^2 statistic and its p-value, and the Akaike Information Criterion (AIC). The χ^2 statistic reveals that adding any of the three surprisal sources significantly improves the fit over and above the covariates, expertise and terminology predictors.

Table C.4 presents the results of the Vuong test. For go-past and total RT, pairwise differences between the reader-aligned surprisal model

Table C.3
Comparison of regression models for the three reading measures in terms of their log-likelihood (LL) and the Akaike Information Criterion (AIC). χ^2 is the test statistic of the likelihood ratio test with significance values indicated as *** ($p < 0.001$), ** ($p < 0.01$), * ($p < 0.05$), no asterisk ($p > 0.05$).

Measure	Surprisal	LL	ΔLL	χ^2	AIC
FPRT	None	−66218.6	–	–	132461.3
	General	−66141.4	77.2	154.4***	132308.9
	Text-aligned	−66135.0	83.6	167.2***	132296.0
	Reader-aligned	−66132.8	85.8	170.6***	132291.6
GP	None	−91014.8	–	–	182053.5
	General	−90904.2	110.6	221.14***	181834.4
	Text-aligned	−90900.3	114.5	228.87***	181826.6
	Reader-aligned	−90860.2	154.6	309.08***	181746.4
TFT	None	−109629.5	–	–	219283.1
	General	−109488.2	141.3	282.7***	219002.4
	Text-aligned	−109484.1	145.4	290.8***	218994.3
	Reader-aligned	−109453.7	175.8	351.6***	218933.4

Table C.4
The Vuong test (non-nested likelihood ratio test), its test statistics and significance values. The asterisk indicates the p-value: *** ($p < 0.001$), ** ($p < 0.01$), * ($p < 0.05$), no asterisk ($p > 0.05$)

Surprisal		m1 fits better than m2 (z)		
m1	m2	FPRT	GP	TFT
Reader-aligned	General	3.19***	6.89***	3.89***
Reader-aligned	Text-aligned	−0.08	9.35***	4.54***
Text-aligned	General	3.16***	1.42	1.34

Table D.5
Reader-aligned surprisal: linear-mixed effects models and their basic linear regression counterparts.

	Variable	LMER				LM			
		β	*SE*	*t*	*p*	*β*	*SE*	*t*	*p*
FPRT	(Intercept)	5.62	0.02	336.16	***	5.62	0.002	2813.88	***
	Length	0.18	0.008	23.99	***	0.18	0.003	72.53	***
	LogFreq	0.008	0.008	1.00		0.005	0.003	1.65	
	Position	−0.02	0.005	−4.49	***	−0.02	0.002	−14.51	***
	Expertise	−0.02	0.002	−10.76	***	−0.02	0.002	−8.74	***
	Terminology	0.04	0.007	5.43	***	0.04	0.002	15.96	***
	Expertise:Terminology	−0.01	0.002	−7.50	***	−0.009	0.002	−5.02	***
	Reader-aligned surprisal	0.08	0.006	13.36	***	0.08	0.002	40.36	***
GP	(Intercept)	5.95	0.02	268.34	***	5.94	0.003	2353.58	***
	Length	0.21	0.009	23.91	***	0.21	0.003	66.26	***
	LogFreq	−0.01	0.009	−1.03		−0.02	0.004	−4.84	***
	Position	−0.01	0.005	−2.06	*	−0.01	0.002	−6.73	***
	Expertise	−0.02	0.002	−8.90	***	−0.02	0.002	−7.88	***
	Terminology	0.07	0.009	7.49	***	0.07	0.003	21.20	***
	Expertise:Terminology	−0.02	0.002	−6.72	***	−0.01	0.002	−5.02	***
	Reader-aligned surprisal	0.12	0.007	18.23	***	0.12	0.003	46.36	***
TFT	(Intercept)	6.23	0.003	192.62	***	6.25	0.002	2543.68	***
	Length	0.26	0.009	29.73	***	0.25	0.003	84.22	***
	LogFreq	−0.05	0.01	−5.44	***	−0.04	0.003	−12.05	***
	Position	−0.05	0.006	−8.49	***	−0.04	0.002	−21.71	***
	Expertise	−0.07	0.002	−29.31	***	−0.07	0.002	−30.07	***
	Terminology	0.06	0.009	6.74	***	0.07	0.003	22.52	***
	Expertise:Terminology	−0.02	0.002	−7.71	***	−0.01	0.002	−6.28	***
	Reader-aligned surprisal	0.13	0.007	19.45	***	0.13	0.002	53.89	***

Table D.6
Text-aligned surprisal: linear-mixed effects models and their basic linear regression counterparts.

	Variable	LMER				LM			
		β	*SE*	*t*	*p*	*β*	*SE*	*t*	*p*
FPRT	(Intercept)	5.62	0.02	33.31	***	5.62	0.002	2814.60	***
	Length	0.18	0.008	23.97	***	0.18	0.003	72.48	***
	LogFreq	0.008	0.008	0.99		0.004	0.003	1.56	
	Position	−0.02	0.005	−4.52	***	−0.02	0.002	−14.68	***
	Expertise	−0.02	0.002	−11.74	***	−0.02	0.002	−9.78	***
	Terminology	0.04	0.008	5.59	***	0.04	0.002	16.38	***
	Expertise:Terminology	−0.02	0.002	−8.28	***	−0.01	0.002	−5.85	***
	Text-aligned surprisal	0.08	0.006	13.24	***	0.08	0.002	40.36	***
GP	(Intercept)	5.95	0.02	268.61	***	5.94	0.002	2352.99	***
	Length	0.22	0.009	24.23	***	0.22	0.003	67.23	***
	LogFreq	−0.01	0.009	−1.21		−0.02	0.004	−5.20	***
	Position	−0.01	0.005	−2.22	*	−0.01	0.002	−7.26	***
	Expertise	−0.03	0.002	−13.36	***	−0.03	0.002	−12.39	***
	Terminology	0.07	0.009	8.20	***	0.07	0.003	23.04	***
	Expertise:Terminology	−0.03	0.002	−10.36	***	−0.02	0.002	−8.71	***
	Text-aligned surprisal	0.11	0.007	15.68	***	0.11	0.002	44.37	***
TFT	(Intercept)	6.233	0.03	192.78	***	6.26	0.002	2544.61	***
	Length	0.27	0.009	29.88	***	0.25	0.003	85.01	***
	LogFreq	−0.05	0.01	−5.42	***	−0.04	0.003	−12.19	***
	Position	−0.05	0.005	−8.61	***	−0.04	0.002	−22.19	***
	Expertise	−0.08	0.002	−34.53	***	−0.08	0.002	−35.27	***
	Terminology	0.07	0.009	7.41	***	0.07	0.003	24.68	***
	Expertise:Terminology	−0.03	0.002	−11.84	***	−0.02	0.002	−10.59	***
	Text-aligned surprisal	0.12	0.007	17.76	***	0.12	0.002	52.95	***

and text-aligned or general surprisal reveal that including reader-aligned surprisal yields the best model fit. For first-pass RT, both reader-aligned and text-aligned surprisals are more beneficial than the general surprisal, but comparing them against each other shows that the predictive power of reader-aligned surprisal does not significantly differ from the predictive power of text-aligned surprisal on first-pass reading times.

Appendix D. Linear mixed-effects models and their basic linear regression counterparts from Study 3

See Tables D.5–D.7.

Table D.7
General surprisal: linear-mixed effects models and their basic linear regression counterparts.

	Variable	LMER				LM			
		β	*SE*	*t*	*p*	*β*	*SE*	*t*	*p*
FPRT	(Intercept)	5.61	0.02	335.98	***	5.62	0.002	2809.51	***
	Length	0.18	0.008	23.93	***	0.18	0.003	72.57	***
	LogFreq	0.004	0.008	0.52		0.001	0.003	0.34	
	Position	−0.02	0.005	−4.06	***	−0.02	0.002	−13.06	***
	Expertise	−0.02	0.002	−11.73	***	−0.02	0.002	−9.74	***
	Terminology	0.04	0.008	5.01	***	0.04	0.002	14.56	***
	Expertise:Terminology	−0.02	0.002	−8.29	***	−0.01	0.002	−5.87	***
	General surprisal	0.08	0.006	12.70	***	0.08	0.002	39.37	***
GP	(Intercept)	5.94	0.02	268.20	***	5.93	0.003	2346.03	***
	Length	0.23	0.009	23.70	***	0.21	0.003	65.88	***
	LogFreq	−0.02	0.009	−1.88		−0.02	0.004	−6.91	***
	Position	−0.008	0.006	−1.51		−0.01	0.002	−4.90	***
	Expertise	−0.03	0.002	−13.33	***	−0.03	0.002	−12.15	***
	Terminology	0.06	0.009	6.49	***	0.06	0.003	18.40	***
	Expertise:Terminology	−0.03	0.002	−10.37	***	−0.02	0.002	−8.64	***
	General surprisal	0.12	0.008	15.47	***	0.12	0.003	43.80	***
TFT	(Intercept)	6.22	0.03	192.44	***	6.25	0.002	2535.53	***
	Length	0.26	0.009	29.23	***	0.25	0.003	83.34	***
	LogFreq	−0.06	0.01	−6.22	***	−0.05	0.003	−14.25	***
	Position	−0.04	0.006	−7.89	***	−0.04	0.002	−19.53	***
	Expertise	−0.08	0.002	−34.54	***	−0.08	0.002	−35.08	***
	Terminology	0.05	0.009	5.48	***	0.06	0.003	19.14	***
	Expertise:Terminology	−0.03	0.002	−11.86	***	−0.02	0.002	−10.55	***
	General surprisal	0.13	0.002	16.49	***	0.12	0.002	52.38	***

Data availability

All data and analysis scripts are made available at https://osf.io/78jdq/.

References

Alabi, J., Mosbach, M., Eyal, M., Klakow, D., & Geva, M. (2024). The hidden space of transformer language adapters. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 6588–6607). Bangkok, Thailand: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.acl-long.356, URL https://aclanthology.org/2024.acl-long.356/.

Andrews, S., & Veldre, A. (2020). Wrapping up sentence comprehension: The role of task demands and individual differences. *Scientific Studies of Reading*, *25*, 123–140, URL https://doi.org/10.1080/10888438.2020.1817028.

Aurnhammer, C., & Frank, S. L. (2018). Comparing gated and simple recurrent neural network architectures as models of human sentence processing. In *Annual meeting of the cognitive science society*. URL https://escholarship.org/uc/item/0br7f339.

Bansal, T., Alzubi, S., Wang, T., Lee, J.-Y., & McCallum, A. (2022). Meta-adapters: Parameter efficient few-shot fine-tuning through meta-learning. In *First conference on automated machine learning (main track)*. URL https://openreview.net/forum?id=BCGNf-prLg5.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48, URL https://doi.org/10.18637/jss.v067.i01.

Beinborn, L., & Pinter, Y. (2023). Analyzing cognitive plausibility of subword tokenization. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 4478–4486). Singapore: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.emnlp-main.272, URL https://aclanthology.org/2023.emnlp-main.272/.

Berzak, Y., & Levy, R. (2023). Eye movement traces of linguistic knowledge in native and non-native reading. *Open Mind*, [ISSN: 2470-2986] *7*, 179–196. http://dx.doi.org/10.1162/opmi_a_00084.

Bird, S., & Loper, E. (2004). NLTK: The natural language toolkit. In *Proceedings of the ACL interactive poster and demonstration sessions* (pp. 214–217). Barcelona, Spain: Association for Computational Linguistics, URL https://aclanthology.org/P04-3031.

Boston, M. F., Hale, J., Kliegl, R., Patil, U., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the potsdam sentence corpus. *Journal of Eye Movement Research*, *2*(1), URL https://doi.org/10.16910/jemr.2.1.1.

Chaffin, R., Morris, R. K., & Seely, R. E. (2001). Learning new word meanings from context: A study of eye movements. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *27 1*, 225–235, URL https://doi.org/10.1037/0278-7393.27.1.225.

Chang, T. A., & Bergen, B. K. (2022). Word acquisition in neural language models. In B. Roark, & A. Nenkova (Eds.), *Transactions of the Association for Computational Linguistics*, *10*, 1–16. http://dx.doi.org/10.1162/tacl_a_00444, URL https://aclanthology.org/2022.tacl-1.1.

Cong, Y., Chersoni, E., Hsu, Y.-y., & Lenci, A. (2023). Are language models sensitive to semantic attraction? A study on surprisal. In A. Palmer, & J. Camacho-collados (Eds.), *Proceedings of the 12th joint conference on lexical and computational semantics (*sEM 2023)* (pp. 141–148). Toronto, Canada: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.starsem-1.13, URL https://aclanthology.org/2023.starsem-1.13.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 8440–8451). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.acl-main.747, URL https://aclanthology.org/2020.acl-main.747.

de Varda, A. G., Gatti, D., Marelli, M., & Günther, F. (2024). Meaning beyond lexicality capturing pseudoword definitions with language models. *Computational Linguistics*, [ISSN: 0891-2017] 1–31. http://dx.doi.org/10.1162/coli_a_00527.

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

de Varda, A., & Marelli, M. (2023). Scaling in cognitive modelling: a multilingual approach to human reading times. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 2: short papers)* (pp. 139–149). Toronto, Canada: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.acl-short.14, URL https://aclanthology.org/2023.acl-short.14.

de Varda, A. G., Marelli, M., & Amenta, S. (2023). Cloze probability, predictability ratings, and computational estimates for 205 english sentences, aligned with existing EEG and reading time data. *Behavior Research Methods*, URL https://doi.org/10.3758/s13428-023-02261-8.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*, 193–210, URL https://www.sciencedirect.com/science/article/pii/S0010027708001741.

Dubey, A., Keller, F., & Sturt, P. (2006). Integrating syntactic priming into an incremental probabilistic parser, with an application to psycholinguistic modeling. In N. Calzolari, C. Cardie, & P. Isabelle (Eds.), *Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the association for computational linguistics* (pp. 417–424). Sydney, Australia: Association for Computational Linguistics, http://dx.doi.org/10.3115/1220175.1220228, URL https://aclanthology.org/P06-1053.

Fine, A. B., Jaeger, T. F., Farmer, T. A., Qian, T., & Paterson, K. B. (2013). Rapid expectation adaptation during syntactic comprehension. *PLoS One*, *8*, URL https://doi.org/10.1371/journal.pone.0077661.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*, 3–71, URL https://www.sciencedirect.com/science/article/pii/0010027788900315.

Fossum, V., & Levy, R. (2012). Sequential vs. Hierarchical syntactic models of human incremental sentence processing. In D. Reitter, & R. Levy (Eds.), *Proceedings of the 3rd workshop on cognitive modeling and computational linguistics (CMCL 2012)* (pp. 61–69). Montréal, Canada: Association for Computational Linguistics, URL https://aclanthology.org/W12-1706.

Frank, S. L., & Aumeistere, A. (2024). An eye-tracking-with-EEG coregistration corpus of narrative sentences. *Language Resources and Evaluation*, *58*, 641–657, URL https://doi.org/10.1007/s10579-023-09684-x.

Frank, S. L., & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, *22*(6), 829–834. http://dx.doi.org/10.1177/0956797611409589, PMID: 21586764.

Frank, S. L., Koppen, M., Noordman, L. G. M., & Vonk, W. (2008). World knowledge in computational models of discourse comprehension. *Discourse Processes*, *45*(6), 429–463, URL https://doi.org/10.1080/01638530802069926.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11, URL https://www.sciencedirect.com/science/article/pii/S0093934X14001515.

Gaskell, M. G., & Ellis, A. W. (2009). Word learning and lexical development across the lifespan. *Philosophical Transactions of the Royal Society, Series B (Biological Sciences)*, *364*, 3607–3615, URL https://doi.org/10.1098/rstb.2009.0213.

Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. In A. Sayeed, C. Jacobs, T. Linzen, & M. van Schijndel (Eds.), *Proceedings of the 8th workshop on cognitive modeling and computational linguistics (CMCL 2018)* (pp. 10–18). Salt Lake City, Utah: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/W18-0102, URL https://aclanthology.org/W18-0102.

Greenberg, C. (2023). Evaluating humanness in language models.

Guest, O., & Martin, A. E. (2021). On logical inference over brains, behaviour, and artificial neural networks. *Computational Brain & Behavior*, *6*, 213–227, URL https://doi.org/10.1007/s42113-022-00166-x.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th ACL* (pp. 8342–8360). Online: ACL, URL https://aclanthology.org/2020.acl-main.740/.

Hahn, M., Futrell, R., Levy, R., & Gibson, E. (2022). A resource-rational model of human processing of recursive linguistic structure. *Proceedings of the National Academy of Sciences*, *119*(43), Article e2122602119. http://dx.doi.org/10.1073/pnas.2122602119, URL https://www.pnas.org/doi/abs/10.1073/pnas.2122602119.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the North American chapter of the association for computational linguistics*. URL https://aclanthology.org/N01-1021.

Haller, P., Bolliger, L., & Jäger, L. (2024). Language models emulate certain cognitive profiles: An investigation of how predictability measures interact with individual differences. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Findings of the association for computational linguistics: ACL 2024* (pp. 7878–7892). Bangkok, Thailand: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.findings-acl.469, URL https://aclanthology.org/2024.findings-acl.469/.

Haller, P., Ding, C., Koncic, I., Reich, D. R., Stegenwallner-Schütz, M., & Jäger, L. A. (2024-11-28). Measurement reliability of individual differences in sentence processing: A cross-methodological reading corpus and Bayesian analysis. URL https://osf.io/preprints/psyarxiv/muv4q_v1.

Hayase, J., Liu, A., Choi, Y., Oh, S., & Smith, N. A. (2024). Data Mixture Inference Attack: BPE Tokenizers Reveal Training Data Compositions. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, & C. Zhang (Eds.), *vol. 37*, *Advances in neural information processing systems* (pp. 8956–8983). Curran Associates, Inc., URL https://proceedings.neurips.cc/paper_files/paper/2024/file/10e6dfea9a673bef4a7b1cb9234891bc-Paper-Conference.pdf.

He, R., Liu, L., Ye, H., Tan, Q., Ding, B., Cheng, L., Low, J., Bing, L., & Si, L. (2021). On the effectiveness of adapter-based tuning for pretrained language model adaptation. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)* (pp. 2208–2222). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.acl-long.172, URL https://aclanthology.org/2021.acl-long.172.

Heilbron, M., van Haren, J., Hagoort, P., & de Lange, F. P. (2023). Lexical processing strongly affects reading times but not skipping during natural reading. *Open Mind : Discoveries in Cognitive Science*, *7*, 757–783, URL https://doi.org/10.1162/opmi_a_00099.

Heister, J., Würzner, K., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, A., & Kliegl, R. (2011). dlexDB – eine lexikalische datenbank für die psychologische und linguistische forschung. *Psychologische Rundschau*, *62*(1), 10–20, URL https://doi.org/10.1026/0033-3042/a000029.

Hollenstein, N., Rotsztejn, J., Troendle, M., Pedroni, A., Zhang, C., & Langer, N. (2018). ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Scientific Data*, *5*, URL https://doi.org/10.1038/sdata.2018.291.

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength natural language processing in python. http://dx.doi.org/10.5281/zenodo.1212303.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., & Gelly, S. (2019). Parameter-efficient transfer learning for nlp. In K. Chaudhuri, & R. Salakhutdinov (Eds.), *Proceedings of Machine Learning Research*: *vol. 97*, *ICML* (pp. 2790–2799). PMLR, URL https://proceedings.mlr.press/v97/houlsby19a/houlsby19a.pdf.

Huang, K.-J., Arehalli, S., Kugemoto, M., Muxica, C., Prasad, G., Dillon, B., & Linzen, T. (2024). Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, [ISSN: 0749-596X] *137*, Article 104510. http://dx.doi.org/10.1016/j.jml.2024.104510, URL https://www.sciencedirect.com/science/article/pii/S0749596X24000135.

Jakobi, D. N., Kern, T., Reich, D. R., Haller, P., & Jäger, L. A. (2025). PoTeC: A German naturalistic eye-tracking-while-reading corpus. *Behavior Research Methods*, *57*(8), 211, URL https://doi.org/10.3758/s13428-024-02536-8.

Jian, Y.-C., & Ko, H. W. (2014). Investigating the effects of background knowledge on Chinese word processing during text reading: Evidence from eye movements. *Journal of Research in Reading*, *37*, URL https://doi.org/10.1111/j.1467-9817.2012.01534.x.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87 4*, 329–354, URL https://doi.org/10.1037/0033-295X.87.4.329.

Kaakinen, J. K., Hyönä, J., & Keenan, J. M. (2003). How prior knowledge, WMC, and relevance of information affect eye fixations in expository text. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *29 3*, 447–457, URL https://doi.org/10.1037/0278-7393.29.3.447.

Kendeou, P., & van den Broek, P. (2007). The effects of prior knowledge and text structure on comprehension processes during reading of scientific texts. *Memory & Cognition*, *35*, 1567–1577, URL https://doi.org/10.3758/BF03193491.

Kendeou, P., Rapp, D. N., & van den Broek, P. (2004). The influence of reader's prior knowledge on text comprehension and learning from text. *Progress in Education*, *13*, 189–209.

Kobayashi, G., Kuribayashi, T., Yokoi, S., & Inui, K. (2023). Transformer language models handle word frequency in prediction head. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the association for computational linguistics: ACL 2023* (pp. 4523–4535). Toronto, Canada: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.findings-acl.276, URL https://aclanthology.org/2023.findings-acl.276.

Krieger, B., Brouwer, H., Aurnhammer, C., & Crocker, M. W. (2025). On the limits of LLM surprisal as a functional explanation of the N400 and P600. *Brain Research*, [ISSN: 0006-8993] *1865*, Article 149841. http://dx.doi.org/10.1016/j.brainres.2025.149841, https://www.sciencedirect.com/science/article/pii/S0006899325004020.

Kuribayashi, T., Oseki, Y., Brassard, A., & Inui, K. (2022). Context limitations make neural language models more human-like. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 10421–10436). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2022.emnlp-main.712, URL https://aclanthology.org/2022.emnlp-main.712.

Kuribayashi, T., Oseki, Y., Ito, T., Yoshida, R., Asahara, M., & Inui, K. (2021). Lower perplexity is not always human-like. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)* (pp. 5203–5217). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.acl-long.405, URL https://aclanthology.org/2021.acl-long.405.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). Lmertest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*, 1–26, URL https://doi.org/10.18637/jss.v082.i13.

Laine, M., Polonyi, T., & Abari, K. (2014). More than words: Fast acquisition and generalization of orthographic regularities during novel word learning in adults. *Journal of Psycholinguistic Research*, [ISSN: 1573-6555] *43*(4), 381–396. http://dx.doi.org/10.1007/s10936-013-9259-1.

Lesci, P., Meister, C., Hofmann, T., Vlachos, A., & Pimentel, T. (2025). The causal effect of merge operations in bottom-up tokenisers. In *Proceedings of the 63rd ACL*. ACL.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*, 1126–1177, URL https://doi.org/10.1016/j.cognition.2007.05.006.

Lowell, R., & Morris, R. K. (2014). Word length effects on novel words: Evidence from eye movements. *Attention, Perception, & Psychophysics*, *76*, 179–189, URL https://doi.org/10.3758/s13414-013-0556-4.

McGrath, S. W., Russin, J., Pavlick, E., & Feiman, R. (2024). How can deep neural networks inform theory in psychological science? *Current Directions in Psychological Science*, *33*(5), 325–333, URL https://doi.org/10.1177/09637214241268098.

Merkle, E. C., You, D., & Preacher, K. J. (2016). Testing nonnested structural equation models. *Psychological Methods*, *21 2*, 151–163, URL https://doi.org/10.1037/met0000038.

Merkx, D., & Frank, S. L. (2021). Human sentence processing: Recurrence or attention? In E. Chersoni, N. Hollenstein, C. Jacobs, Y. Oseki, L. Prévot, & E. Santus (Eds.), *Proceedings of the workshop on cognitive modeling and computational linguistics* (pp. 12–22). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.cmcl-1.2, URL https://aclanthology.org/2021.cmcl-1.2.

Michaelov, J. A., Bardolph, M. D., Van Petten, C. K., Bergen, B. K., & Coulson, S. (2024). Strong prediction: Language model surprisal explains multiple N400 effects. *Neurobiology of Language*, [ISSN: 2641-4368] *5*(1), 107–135. http://dx.doi.org/10.1162/nol_a_00105.

Minixhofer, B. (2020). GerPT2: German large and small versions of GPT2. http://dx.doi.org/10.5281/zenodo.5509984, URL https://github.com/bminixhofer/gerpt2.

van Moort, M. L., Koornneef, A., & van den Broek, P. W. (2020). Differentiating text-based and knowledge-based validation processes during reading: Evidence from eye movements. *Discourse Processes*, *58*, 22–41, URL https://doi.org/10.1080/0163853X.2020.1727683.

Mosbach, M. (2023). Analyzing pre-trained and fine-tuned language models. In Y. Elazar, A. Ettinger, N. Kassner, S. Ruder, & N. A. Smith (Eds.), *Proceedings of the big picture workshop* (pp. 123–134). Singapore: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.bigpicture-1.10, URL https://aclanthology.org/2023.bigpicture-1.10.

Nair, S., & Resnik, P. (2023). Words, subwords, and morphemes: What really matters in the surprisal-reading time relationship? In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the association for computational linguistics: EMNLP 2023* (pp. 11251–11260). Singapore: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.findings-emnlp.752, URL https://aclanthology.org/2023.findings-emnlp.752/.

Nieuwland, M. S., & van Berkum, J. J. A. (2006). When peanuts fall in Love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience*, *18*(7), 1098–1111. http://dx.doi.org/10.1162/jocn.2006.18.7.1098.

Oh, B.-D., Clark, C., & Schuler, W. (2022). Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence*, *5*, URL https://doi.org/10.3389/frai.2022.777963.

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

Oh, B.-D., & Schuler, W. (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*, [ISSN: 2307-387X] *11*, 336–350, URL https://doi.org/10.1162/tacl_a_00548.

Ozuru, Y., Dempsey, K. B., & McNamara, D. S. (2009). Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. *Learning and Instruction*, *19*, 228–242, URL https://doi.org/10.1016/j.learninstruc.2008.04.003.

Paape, D., & Vasishth, S. (2022). Conscious rereading is confirmatory: Evidence from bidirectional self-paced reading. *Glossa Psycholinguistics*, URL https://doi.org/10.5070/G6011182.

Pfeiffer, J., Rücklé, A., Poth, C., Kamath, A., Vulić, I., Ruder, S., Cho, K., & Gurevych, I. (2020). AdapterHub: A framework for adapting transformers. In Q. Liu, & D. Schlangen (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations* (pp. 46–54). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.emnlp-demos.7, URL https://aclanthology.org/2020.emnlp-demos.7.

Radach, R. R., & Kennedy, A. (2013). Eye movements in reading: Some theoretical context. *Quarterly Journal of Experimental Psychology*, *66*, 429–452, URL https://doi.org/10.1080/17470218.2012.750676.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *Technical Report, OpenAI*, URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Rayner, K., & Liversedge, S. P. (2011). Linguistic and cognitive influences on eye movements during reading. In S. P. Liversedge, I. D. Gilchrist, & S. Everling (Eds.), *The oxford handbook of eye movements*. Oxford: University Press, ISBN: 9780199539789.

Ricks, T. R., & Wiley, J. (2009). The influence of domain knowledge on the functional capacity of working memory. *Journal of Memory and Language*, *61*(4), 519–537. http://dx.doi.org/10.1016/j.jml.2009.07.007, URL https://www.sciencedirect.com/science/article/pii/S0749596X0900076X.

van Rooij, I., Guest, O., Adolfi, F., de Haan, R., Kolokolova, A., & Rich, P. (2024). Reclaiming AI as a theoretical tool for cognitive science. *Computational Brain & Behavior*, *7*, 616–636, URL https://doi.org/10.1007/s42113-024-00217-5.

van Schijndel, M., & Linzen, T. (2018). A neural model of adaptation in reading. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 4704–4710). Brussels, Belgium: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D18-1499, URL https://aclanthology.org/D18-1499.

van Schijndel, M., & Linzen, T. (2020). Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science*, *45 6*, Article e12988, URL https://doi.org/10.1111/cogs.12988.

Schurer, T., Opitz, B., & Schubert, T. (2020). Working memory capacity but not prior knowledge impact on readers' attention and text comprehension. *vol. 5*, In *Frontiers in education*. URL https://doi.org/10.3389/feduc.2020.00026.

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. In K. Erk, & N. A. Smith (Eds.), *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 1715–1725). Berlin, Germany: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P16-1162, URL https://aclanthology.org/P16-1162/.

Shain, C. (2024). Word frequency and predictability dissociate in naturalistic reading. *Open Mind*, [ISSN: 2470-2986] *8*, 177–201. http://dx.doi.org/10.1162/opmi_a_00119.

Shain, C., Meister, C., Pimentel, T., Cotterell, R., & Levy, R. (2024). Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, *121*(10), Article e2307876121. http://dx.doi.org/10.1073/pnas.2307876121, URL https://www.pnas.org/doi/abs/10.1073/pnas.2307876121.

Škrjanec, I., Broy, F. Y., & Demberg, V. (2023). Expert-adapted language models improve the fit to reading times. *Procedia Computer Science*, [ISSN: 1877-0509] *225*, 3488–3497. http://dx.doi.org/10.1016/j.procs.2023.10.344, 27th International Conference on Knowledge Based and Intelligent Information and Engineering Sytems (KES 2023). URL https://www.sciencedirect.com/science/article/pii/S1877050923015028.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*, 302–319, URL https://www.sciencedirect.com/science/article/pii/S0010027713000413.

Staub, A. (2015). The effect of lexical predictability on eye movements in reading: Critical review and theoretical interpretation. *Lang. Linguistics Compass*, *9*, 311–327, URL https://doi.org/10.1111/lnc3.12151.

Szewczyk, J. M., & Federmeier, K. D. (2022). Context-based facilitation of semantic access follows both logarithmic and linear functions of stimulus probability. *Journal of Memory and Language*, [ISSN: 0749-596X] *123*, Article 104311. http://dx.doi.org/10.1016/j.jml.2021.104311, URL https://www.sciencedirect.com/science/article/pii/S0749596X21000942.

Tarchi, C. (2010). Reading comprehension of informative texts in secondary school: A focus on direct and indirect effects of reader's prior knowledge. *Learning and Individual Differences*, *20*(5), 415–420. http://dx.doi.org/10.1016/j.lindif.2010.04.002, URL https://www.sciencedirect.com/science/article/pii/S1041608010000373.

Thompson, B., Gwinnup, J., Khayrallah, H., Duh, K., & Koehn, P. (2019). Overcoming catastrophic forgetting during domain adaptation of neural machine translation. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 2062–2068). Minneapolis, Minnesota: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/N19-1209, URL https://aclanthology.org/N19-1209.

Timkey, W., & Linzen, T. (2023). A language model with limited memory capacity captures interference in human sentence processing. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the association for computational linguistics: EMNLP 2023* (pp. 8705–8720). Singapore: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.findings-emnlp.582, URL https://aclanthology.org/2023.findings-emnlp.582.

Troyer, M., & Kutas, M. (2018). Harry Potter and the chamber of what?: The impact of what individuals know on word processing during reading. *Language, Cognition and Neuroscience*, *35*, 641–657, URL https://doi.org/10.1080/23273798.2018.1503309.

Troyer, M., Kutas, M., Batterink, L. J., & McRae, K. (2023). Nuances of knowing: Brain potentials reveal implicit effects of domain knowledge on word processing in the absence of sentence-level knowledge. *Psychophysiology*, *61*(1), Article e14422, URL https://doi.org/10.1111/psyp.14422.

Troyer, M., Urbach, T. P., & Kutas, M. (2020). Lumos!: Electrophysiological tracking of (wizarding) world knowledge use during reading. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *46*(3), 476–486, URL https://doi.org/10.1037/xlm0000737.

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Neural information processing systems* (pp. 6000–6010). URL https://dl.acm.org/doi/10.5555/3295222.3295349.

Venhuizen, N. J., Crocker, M. W., & Brouwer, H. (2019). Expectation-based comprehension: Modeling the interaction of world knowledge and linguistic experience. *Discourse Processes*, *56*(3), 229–255. http://dx.doi.org/10.1080/0163853X.2018.1448677.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, *57*, 307–333, URL https://doi.org/10.2307/1912557.

Warstadt, A., & Bowman, S. R. (2022). What artificial neural networks can tell us about human language acquisition. In *Algebraic structures in natural language*. URL https://doi.org/10.1201/9781003205388.

Weiss, A. F., Kretzschmar, F., Schlesewsky, M., Bornkessel-Schlesewsky, I., & Staub, A. (2018). Comprehension demands modulate re-reading, but not first-pass reading behavior. *Quarterly Journal of Experimental Psychology*, *71*, 198–210, URL https://doi.org/10.1080/17470218.2017.1307862.

Welch, C., Gu, C., Kummerfeld, J. K., Perez-Rosas, V., & Mihalcea, R. (2022). Leveraging similar users for personalized language modeling with limited data. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 1742–1752). Dublin, Ireland: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2022.acl-long.122, URL https://aclanthology.org/2022.acl-long.122.

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., & Levy, R. P. (2020). 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* (pp. 1707–1713). URL https://escholarship.org/uc/item/738338tm.

Wilcox, E., Meister, C., Cotterell, R., & Pimentel, T. (2023). Language model quality correlates with psychometric predictive power in multiple languages. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 7503–7511). Singapore: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.emnlp-main.466, URL https://aclanthology.org/2023.emnlp-main.466.

Wilcox, E. G., Pimentel, T., Meister, C., & Cotterell, R. (2024). An information-theoretic analysis of targeted regressions during reading. *Cognition*, *249*, Article 105765. http://dx.doi.org/10.1016/j.cognition.2024.105765, URL https://www.sciencedirect.com/science/article/pii/S0010027724000519.

Wilcox, E. G., Pimentel, T., Meister, C., Cotterell, R., & Levy, R. P. (2023). Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, *11*, 1451–1470. http://dx.doi.org/10.1162/tacl_a_00612, URL https://aclanthology.org/2023.tacl-1.82/.

Williams, R. S., & Morris, R. K. (2004). Eye movements, word familiarity, and vocabulary acquisition. *European Journal of Cognitive Psychology*, *16*, 312–339, URL https://doi.org/10.1080/09541440340000196.

Yan, S., & Jaeger, T. F. (2020). Expectation adaptation during natural reading. *Language, Cognition and Neuroscience*, *35*(10), 1394–1422. http://dx.doi.org/10.1080/23273798.2020.1784447.

Zhang, Y., Warstadt, A., Li, X., & Bowman, S. R. (2021). When do you need billions of words of pretraining data? In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)* (pp. 1112–1125). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.acl-long.90, URL https://aclanthology.org/2021.acl-long.90.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2021). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, *109*(1), 43–76. http://dx.doi.org/10.1109/JPROC.2020.3004555.