

Surprisal Theory is Tautological (without Rational Grounding)

Ryan Cotterell

ETH Zürich

ryan.cotterell@inf.ethz.ch

Abstract

Surprisal theory holds that the human processing difficulty of a linguistic unit is an affine function of its surprisal in context under some language model. I argue this claim is tautological, i.e., for any non-negative difficulty measure over units in context, there exists a language model whose surprisal is an affine function of it under mild technical conditions. Therefore, because any pattern of difficulty is consistent with *some* language model, without an additional constraint on the language model itself, surprisal theory makes no falsifiable predictions. I further claim that the tautology was obscured by an assumption implicit in two decades of psycholinguistic work—that the language model relevant for studying human sentence processing is the distribution that generated the training corpus. Recent empirical work, however, has undermined this assumption, demonstrating that better corpus models can be *worse* predictors of processing difficulty. I conclude that breaking the tautology requires a rationalist intervention, i.e., the language models used for the study of human sentence processing must be derived from an independently motivated model of the comprehender, which could be based on, for instance, memory constraints or processing goals, and that, crucially, does not depend on the behavioral data surprisal theory is meant to explain.

1 Introduction

A foundational observation in psycholinguistics is that human language comprehension is *incremental*. In particular, comprehenders do not wait until the end of an utterance before beginning to interpret it. Instead, they build a partial interpretation of that utterance unit by unit on the fly (Altmann and Kamide, 1999; Tanenhaus et al., 1995).¹ If comprehension proceeds incrementally, then the total

¹Language processing is not fully incremental. Eye-tracking data contain regressive saccades in which readers often revisit earlier material (Wilcox et al., 2024). Nevertheless, in English and other languages with similar orthography, humans’ eye movements tend to proceed left to right, and,

effort of understanding an utterance $\mathbf{u} = u_1 \dots u_T$ naturally decomposes into a series of per-step costs, one for each unit integrated into the evolving representation. Incrementality thus motivates the central theoretical primitive of this paper—**processing difficulty** $d(u_t \mid \mathbf{u}_{<t}) \geq 0$, a non-negative function that quantifies the effort a comprehender expends to integrate unit u_t given the preceding context $\mathbf{u}_{<t} \stackrel{\text{def}}{=} u_1 \dots u_{t-1}$;² the notation is formally introduced in §2.³ In practice, processing difficulty is operationalized through a variety of behavioral and neural measures—self-paced reading times, eye-tracking fixation durations, and event-related potentials, *inter alia*—and a successful theory of sentence processing must explain its variation across positions and contexts. However, for the exposition of this paper, I abstract away from specific operationalizations with the construct d .

1.1 The Tautology of Surprisal Theory

Surprisal theory (Hale, 2001; Levy, 2008) is a widely adopted information-theoretic framework

correspondingly, the standard modeling assumption in the surprisal literature is that difficulty at position t depends on $\mathbf{u}_{<t}$ alone. I adopt this assumption throughout; extending the framework to account for regressions is an open problem.

²The construct d is agnostic about the source of the processing effort: it subsumes prediction disconfirmation, the cost of integrating u_t into the evolving representation, and any reanalysis of earlier material that u_t triggers. Because reanalysis at position t is initiated by observing u_t in context $\mathbf{u}_{<t}$, its cost remains a function of the pair $(u_t, \mathbf{u}_{<t})$ and is therefore representable by d ; indeed, under Levy’s (2008) resource-reallocation interpretation, reanalysis costs are part of what surprisal is meant to quantify.

³I treat d as a function of the linguistic unit and its context alone, abstracting over reader identity, reader state, and presentation modality. This follows the standard linear mixed-effects regression framework used in psycholinguistics (Smith and Levy, 2013; Goodkind and Bicknell, 2018), which decomposes observed reading times into a surprisal-driven component shared across readers, control predictors (word length, frequency, position), and residual noise. If d were instead reader-specific, the tautology construction in Eq. (2) would produce a different language model for each reader—but surprisal theory posits a single language model shared across readers, which is itself a testable constraint that the present argument does not exploit.

for prediction-based processing (Kuperberg and Jaeger, 2016). It claims that the processing difficulty of a unit in context is an affine function of that unit’s contextual surprisal under some language model q (Shain et al., 2024);⁴ in symbols, surprisal theory corresponds to

$$d(u_t | \mathbf{u}_{<t}) = -a \log \tilde{q}(u_t | \mathbf{u}_{<t}) + b(\mathbf{u}_{<t}), \quad (1)$$

where $a > 0$ is a constant and $b: \Sigma^* \rightarrow \mathbb{R}$ is a context-dependent baseline. This paper argues that, in its current conceptualization, surprisal theory is not falsifiable (Popper, 1959) without constraints that go beyond current practice and, therefore, tautological. Because the theory leaves q as a free parameter, *any* observed pattern of processing difficulty can be accommodated.

A Formal Tautology. To see the emergence of a formal tautology, for any difficulty measure d , one can simply construct a language model whose surprisal is an affine function of it:

$$\tilde{q}_d(u_t | \mathbf{u}_{<t}) \stackrel{\text{def}}{=} \frac{\exp(-d(u_t | \mathbf{u}_{<t}))}{\sum_{u' \in \Sigma_E} \exp(-d(u' | \mathbf{u}_{<t}))}. \quad (2)$$

Under this construction, $-\log \tilde{q}_d(u_t | \mathbf{u}_{<t}) = d(u_t | \mathbf{u}_{<t}) + \log Z(\mathbf{u}_{<t})$, where $Z(\mathbf{u}_{<t})$ is the partition function—so Eq. (1) holds with $a = 1$ and $b(\mathbf{u}_{<t}) = -\log Z(\mathbf{u}_{<t})$.⁵ Under mild conditions, explicated in §3, this construction yields a valid language model.⁶ The consequence is that, without further constraining q , surprisal theory’s claim that difficulty scales with surprisal is vacuous, i.e., it is compatible with every possible choice of d .

The Grounding Problem. Escaping the tautology thus requires two choices that are external to surprisal theory. First, one must fix the family of language models $\mathcal{Q}$ under consideration, which

encodes a structural commitment—for example, one could restrict $\mathcal{Q}$ to the set of all probabilistic context-free grammars (PCFGs; Booth and Thompson, 1973). Second, one must give a criterion by which the language model $q \in \mathcal{Q}$ is selected, e.g., whether it is fit to a corpus, derived from cognitive principles, or optimized against behavioral data. I call the need to make both choices surprisal theory’s **grounding problem**. Hale (2001) confronted the first choice directly. Invoking his Principle 1, which he terms strong competence (Chomsky, 1965), he restricts $\mathcal{Q}$ to the set of PCFGs on rationalist grounds—specifically, he contends that “the human sentence processing mechanism directly uses rules of grammar in its operation.” I examine the second choice further in §1.3.

The Drift. One possible rebuttal of my claimed tautology is that proponents of surprisal theory have never actually claimed that just *any* language model will suffice. On this point, I agree; this is not what surprisal theory’s proponents are *claiming*. Instead, my critique focuses on the *practice* exhibited in many recent papers on surprisal theory. The problem, as I see it, is that subsequent empirical work has progressively *drifted* from commitments such as Hale’s (2001). For instance, practitioners have moved from using PCFGs to estimate surprisal to using recurrent neural networks (Hochreiter and Schmidhuber, 1997) and transformers (Vaswani et al., 2017)—whichever architecture is currently in vogue—and most papers on the subject adopt model families whose cognitive relevance the authors do not even argue for.

The Cognitivist Assumption. A second possible rebuttal appeals to Marr’s (1982) levels of analysis: the *computational* level specifies what problem a system solves and why, the *algorithmic* level specifies the representations and procedures by which it solves it, and the *implementational* level specifies how those procedures are physically realized. Surprisal theory was, from its inception, presented as a computational-level theory, and one could argue that the cognitive grounding of $\mathcal{Q}$ is an algorithmic-level detail—so that the drift just described is irrelevant to a claim that was never algorithmic in the first place.⁷ Even granting this,

⁴Shain et al. (2024) provide large-scale evidence that the relationship between word predictability and reading time is linear in log-probability—i.e., that the affine form in Eq. (1) is a better fit than alternatives such as a logarithmic or power-law relationship in raw probability. The affine formulation is important for my tautology argument: the construction in Eq. (2) yields $a = 1$ and a context-dependent intercept $b = -\log Z(\mathbf{u}_{<t})$, which is exactly the affine form. A strictly proportional formulation ($b = 0$) would leave a gap, since the partition function varies with context.

⁵In psycholinguistic practice, context-dependent baselines are absorbed into control predictors in the regression model. The affine formulation with context-dependent b is therefore the natural statement of the theory.

⁶The tautology argument applies more broadly to any theory that defines processing difficulty purely as a function of the language model’s probability distribution.

⁷Bux6-Lugo et al. (2026) adduce a similar argument about surprisal theory in terms of Marr’s (1982) levels. Like me, they consider surprisal theory to be a computational-level account. However, they then argue that the specific algorithmic-level commitments, e.g., the choice of language model used to

one must still say what q is meant to approximate. My answer is a theoretical construct p_H , the comprehender’s internal language model; I call the assumption that q approximates p_H the **cognitivist assumption**. Levy (2008) comes closest to making the cognitivist assumption explicit, writing “the cognitive entity of primary interest is the resulting probabilistic [unit] model alone” (p. 1138), and that the choice of grammar “is not a commitment to any particular grammatical formalism as the backbone of sentence comprehension” but rather “a formal means of estimating what expectations about upcoming [units] in a sentence implicitly arise” (p. 1138). Under the cognitivist assumption, surprisal theory reduces to the claim that

$$d(u_t \mid \mathbf{u}_{<t}) = -a \log \vec{p}_H(u_t \mid \mathbf{u}_{<t}) + b(\mathbf{u}_{<t}). \quad (3)$$

But the cognitivist assumption does not, by itself, solve the grounding problem: without an independent characterization of p_H that is *not* grounded in the behavioral data that surprisal theory seeks to explain, any pattern of difficulty can be explained by positing a p_H whose surprisal matches it. It merely relocates the problem; it does not remove it.

1.2 An Analogy from Evolutionary Biology

Tautologies are not new in science. An instructive parallel comes from evolutionary biology. Darwin’s (1859) principle of natural selection—that heritable traits conferring reproductive advantage spread through a population—is distilled in the slogan *survival of the fittest*. But if fitness simply means reproductive success, the slogan reduces to “those who reproduce most, reproduce most”—a tautology. The observation has a long history (Butler, 1882; Sinnott et al., 1958; Waddington, 1960) and was given its most sustained treatment by Peters (1976); Popper famously concluded that natural selection was a “metaphysical research programme” rather than a testable theory (Popper, 1974). The criticism was not that evolution was

compute surprisal, sneak into the analysis, resulting in a con-found. This argument has some substance, but I worry that their claim proves too much. Indeed, *any* measurement of a computational theory would fall victim to their concern. In fairness, the authors pre-empt this objection (p. 25), arguing that the problem lies specifically in using “black box” models as measurement tools to test a computational theory. On my reading, Buxó-Lugo et al. (2026) fail to offer a definition of blackboxhood, or even a criterion by which one could claim that a transformer language model is blackbox, but a PCFG is not. Moreover, I remain unconvinced that an ontological distinction between blackbox and non-blackbox models can be pinned down. My argument, in contrast, concerns a tautology present purely at the computational level of analysis.

false, but that its central explanatory construct was defined in terms of the very phenomenon it was meant to explain.

The resolution came through what Mills and Beatty (1979) called the propensity interpretation of fitness; see also Brandon (1978). On this view, fitness is not a measure of actual reproductive output but a **propensity**—a causal tendency, grounded in an organism’s morphology, physiology, and behavior in a given environment, to survive. Under the propensity interpretation, natural selection makes a falsifiable claim: organisms with higher propensities reproduce more than those with lower ones. The key structural feature of the propensity interpretation—and the one I will exploit below—is that it grounds an explanatory construct in physical or structural properties of the organism itself and how it relates to its environment. Crucially, this grounding is *independent* of the outcome the construct is meant to explain—reproductive success. A propensity interpretation of surprisal theory would require an analogous move: grounding the choice of q in independently observable properties of the comprehender—working memory capacity, parsing strategy, attention allocation policy—rather than in the reading-time data the theory is meant to predict. I discuss concrete proposals along these lines in §5.

1.3 How Surprisal Theory Lost Its Propensity

While rarely stated explicitly, I contend that surprisal theory *did* historically have a propensity interpretation, which I explicate below. Let p_C denote the data-generating language model of the corpus—the distribution over strings that produced the texts used to estimate q . Just as fitness under the propensity interpretation is grounded in how the organism interacts with its environment—observable physical properties *independent* of reproductive outcomes— p_C is grounded in how humans produce text in the wild, observable independently of any processing difficulty measurements.⁸

⁸One might object that the independence is not absolute, i.e., the same cognitive system that produces text also processes it, so p_C and processing difficulty are not causally unrelated. The dependence may run deeper than mere causal correlation: influential theories of language processing posit shared representations for production and comprehension (Pickering and Garrod, 2004; Dell, 1986), in which case p_C is not independent of p_H but rather a noisy reflection of it. If production and comprehension share the same underlying language model, the corpus assumption and the cognitivist assumption would collapse into one—an identification that, if correct, would simplify the grounding problem considerably.

To state it more concisely, I call the assumption that I contend has anchored two decades of empirical work the **corpus assumption**. Formally, the corpus assumption holds that $q \in \mathcal{Q}$ is chosen such that $q = p_C$. Or, when equality is not achievable, one should choose q so as to minimize $D_{\text{KL}}(p_C \parallel q)$. Under the corpus assumption, surprisal theory makes the following claim:

$$d(u_t \mid \mathbf{u}_{<t}) = -a \log \vec{p}_C(u_t \mid \mathbf{u}_{<t}) + b(\mathbf{u}_{<t}). \quad (4)$$

One wrinkle, of course, is that p_C is not known. However, it can be approximated by estimating a language model from a corpus of texts assumed to have been sampled from p_C . The corpus assumption led to two concrete payoffs. First, it made finding a q to experiment with in practice approachable: one can estimate a language model q from a corpus and treat it as an approximation to p_C . Second, under maximum-likelihood estimation, the corpus assumption entails what I call the **scaling implication**—as estimation improves, a model q should asymptotically yield surprisal values that better correlate with processing difficulty. This implication is a direct consequence of the corpus assumption, as I formally argue in §4, and forces surprisal theory to make a falsifiable prediction.

As a caveat, I note that, in practice, the corpus assumption has seldom been enforced in pure form. Studies routinely compare many language models by their fit to behavioral data and report the best-fitting configurations (e.g., Goodkind and Bicknell, 2018; Wilcox et al., 2023; Oh and Schuler, 2023), so that subsequent work, in adopting those models, implicitly conditions the choice of q on the very data the theory is meant to explain; hyperparameter and checkpoint selection also introduce a logical dependence, albeit a mild one. At the extreme, Kiegl et al. (2024) fine-tune a language model on reading times directly; I return to the status of this strategy in §4.⁹

The point, for present purposes, is that p_C can be *estimated* without recourse to reading-time data, which is sufficient for the propensity move to go through. I concede, however, that the relationship between production and comprehension distributions deserves more careful treatment than I am able to give it here.

⁹In the same vein, Yoshida et al. (2026) fine-tune neural language models on garden-path reading times, yielding an “existence proof” of a language model whose surprisal reproduces garden-path effects that off-the-shelf models notoriously underestimate—and they note that such constructibility may render surprisal theory “unfalsifiable in practice.” Their empirical construction is the practical counterpart of the formal one in §3; notably, however, their fine-tuned models still fail

Initially, results suggested that contextual surprisal derived from larger language models did in fact correlate more strongly with a variety of behavioral measurements, e.g., self-paced reading times (Smith and Levy, 2013; Wilcox et al., 2023), eye-tracking measures (Boston et al., 2008; Demberg and Keller, 2008), and neural signals (Frank et al., 2015; Giulianelli et al., 2024) across languages and model architectures. Recently, however, a growing body of evidence appears to falsify the scaling implication—and, with it, the corpus assumption specifically. Specifically, Oh and Schuler (2023) reported that the correlation between surprisal and reading times improves as transformers grow larger, but only up to a point—beyond which it *degrades* as models become better approximations to p_C .¹⁰ This finding has since been replicated and extended. Oh et al. (2024) confirmed the inverse relationship across four language model families and four reading corpora. Lin and Schuler (2025) demonstrate that the same inverse trend appears for fMRI-based neural measures, extending the finding beyond reading-time data. Most definitively, Oh et al. (2025) claim to have ruled out data leakage as an explanation, showing that the inverse relationship persists even with models trained on data from which reading-time corpora have been removed. Assuming these findings are correct, then the corpus assumption is no longer a viable foothold for surprisal theory.¹¹

2 Formal Preliminaries

Alphabets, strings, and language models. An **alphabet** Σ is a finite, non-empty set whose elements I call **units**—for instance, words, characters,

to reproduce memory-based effects such as the subject–object relative-clause asymmetry, showing that the family reachable by fine-tuning in practice is not yet universal—a contingent bound on flexibility that, as the formal argument shows, cannot be relied upon in principle.

¹⁰Kuribayashi et al. (2025) argue that this degradation is an artifact of evaluating only final-layer representations: middle layers of larger models retain strong alignment with human reading times. Relatedly, Tsipidi et al. (2026) find that early-layer representations outperform surprisal itself in predicting early-pass reading measures for English across two datasets. However, their results were less consistent in other languages, and these findings do not rescue the scaling implication, which concerns the model’s closeness to p_C .

¹¹If one considers a corpus of conversational speech rather than web text, the scaling implication might still hold for models trained on such corpora. Distinguishing “no data-generating distribution p_C yields the right q ” from “the right p_C has simply not been identified” is an open empirical question that the formal argument of this paper does not resolve.

or subword tokens.^{12,13} The **Kleene closure** Σ^* is the set of all finite strings over Σ , including the empty string ε . I denote individual units by u (or u_t when indexed by position), strings by $\mathbf{u} = u_1 \dots u_T$ where $|\mathbf{u}| = T$ denotes the length, and prefixes of a string $\mathbf{u}$ are denoted by $\mathbf{u}_{< t} = u_1 \dots u_{t-1}$. I augment Σ with a distinguished end-of-sequence symbol $\text{EOS} \notin \Sigma$ and write $\Sigma_E \stackrel{\text{def}}{=} \Sigma \cup \{\text{EOS}\}$ for the EOS-augmented alphabet. A **language model** p is a probability distribution over Σ^* , i.e., $\sum_{\mathbf{u} \in \Sigma^*} p(\mathbf{u}) = 1$. I use p as a generic language model throughout this section; the specific language models q , p_H , and p_C introduced in §1 are all instances of p . I say p has finite expected length if $\mathbb{E}_{\mathbf{u} \sim p}[|\mathbf{u}|] < \infty$. The **prefix probability** of a string $\mathbf{u}$ under p is $\vec{p}(\mathbf{u}) \stackrel{\text{def}}{=} \sum_{\mathbf{u}' \in \Sigma^*} p(\mathbf{u}\mathbf{u}')$, i.e., the probability that a string sampled from p begins with $\mathbf{u}$.

From language models to conditionals. For each prefix $\mathbf{u}_{< t} \in \Sigma^*$, the language model induces a conditional distribution over the extended alphabet Σ_E , given by $\vec{p}(u_t | \mathbf{u}_{< t}) = \vec{p}(\mathbf{u}_{< t}u_t) / \vec{p}(\mathbf{u}_{< t})$ where $u_t \in \Sigma_E$. Because $\vec{p}(\mathbf{u}_{< t}u_t)$ is a sub-sum of $\vec{p}(\mathbf{u}_{< t})$, the numerator vanishes whenever the denominator does. I adopt the convention $0/0 \stackrel{\text{def}}{=} 0$, under which the conditional is defined for every prefix $\mathbf{u}_{< t}$ and is identically zero on prefixes of probability zero. For $u_t = \text{EOS}$, I adopt the convention $\vec{p}(\mathbf{u}_{< t} \text{EOS}) \stackrel{\text{def}}{=} p(\mathbf{u}_{< t})$ —the probability that the string ends exactly at $\mathbf{u}_{< t}$ —since the prefix probability is otherwise defined only for strings in Σ^* . This conditional specifies the probability of each next unit $u_t \in \Sigma_E$ given the context—including the probability of ending the string by generating EOS. Thus, every language model p decomposes as a product of these conditionals as

$$p(\mathbf{u}) = \vec{p}(\text{EOS} | \mathbf{u}) \prod_{t=1}^T \vec{p}(u_t | \mathbf{u}_{< t}), \quad (5)$$

where $\mathbf{u} = u_1 \dots u_T$ is a string of length T .¹⁴ I call this a language model’s **autoregressive factorization**. Furthermore, Eq. (5) allows one to view p as a collection of conditional distributions $\{\vec{p}(\cdot | \mathbf{u})\}_{\mathbf{u} \in \Sigma^*}$, each over Σ_E .

¹²See Kiegele et al. (2026) for a careful treatment of the choice of unit inventory in surprisal-based analyses.

¹³Although Σ is finite, it can encode structured objects with unbounded complexity through serialization into strings. Snæbjarnarson et al. (2026) show how infinite-state languages can be represented over a finite alphabet via such encodings.

¹⁴Note that T is itself a random variable in this formulation.

From conditionals to language models. The converse direction is more delicate. A **conditional collection** is any family $\{\vec{p}(\cdot | \mathbf{u})\}_{\mathbf{u} \in \Sigma^*}$ of distributions over Σ_E , one for each prefix, where $\vec{p}$ is not assumed to arise from a language model. A conditional collection can be used to compute the probability of a string by plugging the appropriate conditionals into Eq. (5). However, the resulting measure over Σ^* may fail to sum to 1: probability mass can “leak” onto infinite sequences that never produce EOS. A conditional collection for which no such leakage occurs, i.e., for which $\sum_{\mathbf{u} \in \Sigma^*} p(\mathbf{u}) = 1$, is called **tight**. A characterization of tightness is given by Du et al. (2023, Theorem 4.7).

3 Technical Conditions for the Tautology

In §1, I argued that any non-negative difficulty measure can be converted into a language model whose surprisal reproduces it; see Eq. (2). I now make this precise. Given a difficulty measure $d: \Sigma_E \times \Sigma^* \rightarrow \mathbb{R}_{\geq 0}$, the construction in Eq. (2) defines a conditional collection q_d over Σ_E :

$$\vec{q}_d(u_t | \mathbf{u}_{< t}) \stackrel{\text{def}}{=} \frac{\exp(-d(u_t | \mathbf{u}_{< t}))}{\sum_{u' \in \Sigma_E} \exp(-d(u' | \mathbf{u}_{< t}))}. \quad (6)$$

The technical subtlety is that q_d as defined in Eq. (6) is not guaranteed to be a language model, as discussed in §2. One could translate Du et al.’s (2023) exact characterization of the tightness of conditional collections directly into the notation of d , but the resulting expression is cumbersome. Instead, I give a simpler sufficient condition purely in terms of d , which admits a more lucid discussion about cognitive plausibility.

Proposition 1. *Let $d: \Sigma_E \times \Sigma^* \rightarrow \mathbb{R}_{\geq 0}$ be a non-negative difficulty measure, and let q_d be defined by Eq. (6). If there exists a function $f: \mathbb{Z}_{>0} \rightarrow \mathbb{R}_{\geq 0}$ such that*

$$d(\text{EOS} | \mathbf{u}) \leq f(|\mathbf{u}| + 1) \quad \text{for all } \mathbf{u} \in \Sigma^* \quad (7)$$

and $\sum_{t=1}^{\infty} \exp(-f(t)) = \infty$, then (a) q_d is tight (see §2), i.e., $\sum_{\mathbf{u} \in \Sigma^*} q_d(\mathbf{u}) = 1$; and (b) if moreover f is constant—i.e., there exists $C > 0$ such that $d(\text{EOS} | \mathbf{u}) \leq C$ for all $\mathbf{u} \in \Sigma^*$ —then q_d has finite expected length: $\mathbb{E}_{\mathbf{u} \sim q_d}[|\mathbf{u}|] < \infty$.

The proof is deferred to App. B.1.

Cognitive plausibility. I now discuss the cognitive plausibility of part (b) of Proposition 1. In words, it requires that $d(\text{EOS} | \mathbf{u})$ —the difficulty of wrapping up the sentence after a prefix $\mathbf{u}$ —be

uniformly bounded. Reading times at sentence-final words do exhibit a well-documented wrap-up cost (Smith and Levy, 2013; Meister et al., 2022), but this cost is a fixed penalty associated with sentence-boundary integration and does not appear to scale with the length of the preceding context. The assumption that d is bounded—which I adopt throughout—therefore satisfies the condition for any realistic difficulty measure.¹⁵ Part (b) is needed as a technical condition in §4, where the scaling implication requires q_d to have finite expected length.

4 The Scaling Implication

The corpus assumption (§1.3) identifies the target language model with the data-generating distribution: $q = p_C$. A direct consequence is that one should choose as expressive a family $\mathcal{Q}$ as possible, and, moreover, the richer $\mathcal{Q}$ is, the better the estimator can approximate p_C , and the closer its surprisal values should be to the true surprisal under p_C .

4.1 Measuring Psychometric Fitness

To formalize the scaling implication, I require a measure of how well a language model’s surprisal values predict processing difficulty. In practice, researchers operationalize this fit by regressing behavioral measures (reading times, fixation durations, or ERP amplitudes) against surprisal values and reporting a goodness-of-fit statistic—typically delta log-likelihood (Goodkind and Bicknell, 2018; Wilcox et al., 2023), R^2 (Giulianelli et al., 2024), or mean-squared error (Kiegeland et al., 2024). I abstract over these choices with a general definition.

Definition 1 (Fitness measure). *Let p be a language model with finite expected length (§2) and let $g: \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}$ be a bounded, continuous function.¹⁶ The (p, g) -fitness measure is defined as*

$$M(q) \stackrel{\text{def}}{=} \sum_{u \in \Sigma_E} \sum_{\mathbf{u} \in \Sigma^*} \tilde{p}(\mathbf{u}u) g(d(u | \mathbf{u}), -\log \tilde{q}(u | \mathbf{u})), \quad (8)$$

¹⁵Concretely, if $d(u | \mathbf{u}) \leq C$ for all u and $\mathbf{u}$, then the constructed model satisfies $\tilde{q}_d(u | \mathbf{u}) \geq \exp(-C)/|\Sigma_E|$, so q_d itself meets the assumption in Definition 1.

¹⁶I assume throughout that per-unit probabilities are uniformly bounded away from zero: there exists $\eta > 0$ such that $\tilde{q}(u | \mathbf{u}) \geq \eta$ for every $q \in \mathcal{Q}$, $u \in \Sigma_E$, and $\mathbf{u} \in \Sigma^*$; necessarily $\eta \leq 1/|\Sigma_E|$, since the conditional probabilities sum to one. Softmax-based models whose parameters range over a compact set satisfy this assumption. Two consequences are used in what follows: first, surprisal values lie in the compact interval $[0, \log(1/\eta)]$; second, by the geometric argument in the proof of Proposition 1, part (b), every $q \in \mathcal{Q}$ is tight with $\mathbb{E}_{\mathbf{u} \sim q}[|\mathbf{u}|] \leq (1 - \eta)/\eta$.

where $\tilde{p}(\mathbf{u}u)$ is the prefix probability of $\mathbf{u}u$ under p , as defined in §2, with the convention $\tilde{p}(\mathbf{u} \text{ EOS}) \stackrel{\text{def}}{=} p(\mathbf{u})$.

For delta log-likelihood, R^2 , and mean-squared error, boundedness of g can be imposed without loss of generality due to the fact that the per-unit probabilities are bounded away from zero. See footnote 16: together with the boundedness of d (§3), it confines the arguments of g to the compact rectangle $[0, C] \times [0, \log(1/\eta)]$ with $\eta > 0$, and M depends on g only through its values there; any continuous g may therefore be truncated outside this rectangle without changing M , so the boundedness requirement sacrifices no generality.

Lemma 1 (Continuity of M). *Suppose p has finite expected length and g is continuous and bounded. Then $M: \mathcal{Q} \rightarrow \mathbb{R}$ is continuous.*

The proof is deferred to App. B.2.

4.2 Estimation

Given a (p_C, g) -fitness measure M , one seeks the language model $q^* = \arg \max_{q \in \mathcal{Q}} M(q)$ that best predicts processing difficulty.¹⁷ The choice of $\mathcal{Q}$ is critical. If $\mathcal{Q}$ is the set of all language models—all distributions over Σ^* —then by Proposition 1, q^* can perfectly accommodate any observed difficulty pattern, and surprisal theory has no grounding. The smaller $\mathcal{Q}$ is, the more constrained and falsifiable surprisal theory becomes. Under the corpus assumption, however, one need not justify $\mathcal{Q}$ on cognitive grounds at all: the goal is simply to approximate p_C as well as possible, so in principle one would choose the most expressive family available.

Because one does not have access to p_C , a standard method for selecting $q \in \mathcal{Q}$ under the corpus assumption is **maximum-likelihood estimation** (MLE).¹⁸ Given a corpus $\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(N)} \stackrel{\text{i.i.d.}}{\sim} p_C$ of

¹⁷As discussed in §1.3, Kiegeland et al. (2024) pursued exactly this strategy; as they themselves note, it raises the question of what, if anything, the optimized model explains.

¹⁸In practice, neural language models are trained by stochastic gradient descent (SGD) on the cross-entropy loss, which is the empirical analogue of the KL objective but does not guarantee convergence to the global MLE. The loss landscape of transformer language models is non-convex with many local minima and saddle points, and SGD may converge to a spurious local optimum rather than the global KL-minimizer. The scaling implication, as stated, is therefore a claim about the *population-level* optimum within $\mathcal{Q}$ —the best model that *could* be found—not about the output of any particular training run. Additional theory, perhaps drawing on the empirical observation that overparameterized networks tend to find good minima at scale, may be necessary to bridge this gap.

N strings, the MLE is

$$q_N \stackrel{\text{def}}{=} \arg \max_{q \in \mathcal{Q}} \frac{1}{N} \sum_{n=1}^N \log q(\mathbf{u}^{(n)}). \quad (9)$$

Under standard regularity conditions—compactness of $\mathcal{Q}$ and a uniform integrability condition on $\log q$, both made precise in Theorem 1 of App. A—the estimator converges almost surely to the KL projection of p_C onto $\mathcal{Q}$:

$$D_{\text{KL}}(p_C \parallel q_N) \xrightarrow[N \rightarrow \infty]{\text{a.s.}} \inf_{q \in \mathcal{Q}} D_{\text{KL}}(p_C \parallel q). \quad (10)$$

Let $q^* \stackrel{\text{def}}{=} \arg \min_{q \in \mathcal{Q}} D_{\text{KL}}(p_C \parallel q)$ denote the KL projection of p_C onto $\mathcal{Q}$. When $\mathcal{Q}$ is well-specified—i.e., $p_C \in \mathcal{Q}$ —the right-hand side of Eq. (10) is zero and $q^* = p_C$. Under misspecification, q^* is the member of $\mathcal{Q}$ closest to p_C in KL divergence. In either case, the following holds.¹⁹

Proposition 2 (Scaling implication). *Assume $\mathcal{Q}^*$ is a singleton $\{q^*\}$, i.e., the KL-minimizer is unique. Under the regularity conditions of Theorem 1, it holds that*

$$M(q_N) \xrightarrow[N \rightarrow \infty]{p} M(q^*). \quad (11)$$

In particular, for any $\epsilon > 0$ and $\delta > 0$, there exists N_0 such that for all $N \geq N_0$, it holds that $\Pr(|M(q_N) - M(q^*)| < \epsilon) > 1 - \delta$.

The proof is deferred to App. B.3.

Proposition 2 establishes that fitness *converges* to $M(q^*)$ as the estimator improves—but it does not predict that fitness increases monotonically with more data or larger models, which is the informal reading of “scaling.” Thus, the convergence result itself is not what is falsified. What the corpus assumption additionally predicts—and what Oh and Schuler’s (2023) findings contradict—is that p_C *maximizes* M , i.e., that the true corpus distribution is the best predictor of processing difficulty. If M first increases and then decreases as models approach p_C , then an imperfect approximation to p_C is a better predictor than p_C itself—a reversal that misspecification alone cannot produce.²⁰

¹⁹When identifiability fails, the conclusion of Proposition 2 still holds provided M takes the same value on all elements of $\mathcal{Q}^*$, since every limit point of $\{q_N\}$ lies in $\mathcal{Q}^*$ by Theorem 1.

²⁰Strictly, this inference requires that the largest models approach p_C itself, not merely the KL projection $q^* \neq p_C$: under misspecification, a rise-and-fall of M along the path to q^* is compatible with p_C maximizing M . The expressiveness of modern architectures makes approximate well-specification plausible, but it is an assumption, not a theorem.

Even before this empirical falsification, the corpus assumption deserved more scrutiny: p_C is determined by who wrote the texts, when, and for what purpose—a model estimated from Wall Street Journal prose (Marcus et al., 1993) converges to a different distribution than one estimated from Reddit (cf. Radford et al., 2019), encoding corpus-specific statistics (genre, topic, editorial style) with no clear role in accounts of syntactic processing. The grounding problem is not solved but deferred: from “which q ?” to “which corpus?”

5 The Search for a Rationalist Grounding

If one accepts the corpus assumption as falsified and the early rationalist commitments of Hale (2001) abandoned with the move to neural models, the question becomes whether surprisal theory can be rescued by choosing $\mathcal{Q}$ on cognitive grounds. What I call the **rationalist principle** is the commitment to grounding $\mathcal{Q}$ and q in aspects of human cognition motivated independently of the behavioral data surprisal theory is meant to explain.²¹ The clearest examples are constraints on the comprehender’s processing architecture—working memory capacity, incremental processing—established through non-reading-time paradigms.

5.1 Lossy-Context Surprisal

Futrell et al. (2020) propose replacing the full context $\mathbf{u}_{<t}$ with a lossy compression $m(\mathbf{u}_{<t})$, where m is a deletion noise process reflecting memory decay. If m is derived from independently motivated assumptions about human memory rather than fit to reading-time data, this move can return propensity to surprisal theory. Hahn et al. (2022) scale this framework by parameterizing m as a resource-rational retention policy optimized to minimize expected surprisal under a fixed resource budget—a rational analysis of the comprehender’s processing problem, not a fit to behavioral data. They confirm the model’s predictions in three experiments on center-embedded relative clauses, producing fine-grained predictions that differ sharply from standard surprisal and that are borne out empirically. However, the grounding is partial: m is derived from first principles, but the underlying q (GPT-2) is not. A complementary route is to build the memory limitation into the architecture itself: Timkey

²¹I use “rationalist” loosely. Specifically, I was inspired by the nativist tradition of Chomsky (1966).

and Linzen (2023) construct a recurrent language model with a single self-attention head whose limited retrieval capacity mirrors cue-based retrieval theories of working memory, and show that it captures semantic and syntactic interference effects observed in human sentence processing—thereby grounding the choice of Q itself, not merely m , in independently motivated memory constraints. McCurdy and Hahn (2024) and De Varda and Marelli (2024) extend this line of work, showing that cognitively justified context degradation improves alignment with human reading times. The promise of lossy-context surprisal lies in the independent derivation of m : a cognitively motivated memory function can break the tautology in a way that an empirically optimized one—which would merely shift the circularity from q to m —cannot.

5.2 Developmentally Plausible Corpora

A different strategy is the BabyLM approach (Warstadt et al., 2023; Hu et al., 2024; Charpentier et al., 2025), which restricts training data to developmentally plausible corpora, e.g., drawing on the CHILDES corpus (MacWhinney, 2000), thereby grounding the corpus choice in language acquisition research rather than reading-time correlations. Two pitfalls remain. First, with a developmentally realistic amount of input, the estimator lacks an asymptotic interpretation, and the resulting model is dominated by its inductive bias (architecture, optimizer, initialization). Unless that bias is independently justified on cognitive grounds, the grounding problem reappears.²² Second, this approach conflates “what children hear” with “how adult humans process language.” One may be able to resolve this conflation satisfactorily—for instance, by arguing that adult processing reflects the statistics of childhood input—but the argument requires independent support, of which I am unaware.

6 Anticipated Objections

I close by addressing two objections that a committed proponent of surprisal theory might raise.

Bounded Rationality. First, one might respond that the corpus assumption was never unprincipled.

²²One might object that regularization implicitly constrains Q in cognitively relevant ways. If regularization is what makes intermediate-sized models predict reading times better, that would constitute evidence for a simplicity constraint on p_H —a promising direction, but one requiring independent theoretical motivation.

Under rational analysis (Anderson, 1990), the comprehender’s expectations should approximate the statistics of the environment because optimal prediction demands it, so the corpus assumption is itself derived from an independently motivated commitment. One could argue that the degradation documented by Oh and Schuler (2023) is thus evidence of *bounded* rationality—not a collapse of grounding. I welcome this rebuttal because it concedes the thesis. Indeed, I contend that it is just a reframing of my argument. It grounds q in a rationalist commitment, i.e., optimal prediction subject to resource constraints, that is external to the behavioral data. Moreover, the appeal to bounded rationality forces one to enumerate those resource constraints. If the bounds are left unspecified, any departure from optimal prediction can be attributed to *some* bound after the fact, and the tautology reappears with the bound, rather than q , as the free parameter. Enumerating the constraints on independent grounds, e.g., the memory limitations that underlie lossy-context surprisal (§5), is precisely what the rationalist principle demands.

Auxiliary Hypotheses. Second, one might object, following Lakatos (1970), that no theory is tested in isolation. Instead, it is the case that the core commitments of a theory are examined jointly with a number of auxiliary hypotheses. Newton’s second law is a classic illustration of this point:

$$F = ma. \quad (12)$$

Here, F denotes a force acting on a body, m the body’s mass, and a the resulting acceleration. If one *defines* a force as whatever F makes the equation hold, then, taken on its own, Eq. (12) is unfalsifiable. Eq. (12) acquires empirical content only once the forces are specified independently, e.g., as with universal gravitation or Hooke’s law. Such force laws are fixed in advance, drawn from a small family, and tested on phenomena to which they were not fitted; the language model is surprisal theory’s counterpart of the force law. Although Newtonian mechanics is thus open to a seemingly similar charge of tautology, no one concludes that it is tautological as a whole. On Lakatos’s (1970) account, this is because a theory is appraised not by the falsifiability of any single prediction but by whether it is *progressive*, i.e., whether its adjustments continue to predict novel facts about the world, or whether it is *degenerating*, i.e., whether they merely accommodate anomalies after the fact.

Newtonian mechanics has repeatedly met this standard, e.g., the existence and location of Neptune were inferred from anomalies in the orbit of Uranus before the planet was identified (Grosser, 1962). Crucially, the same strategy could, and did, fail: a further planet, Vulcan, was posited to explain anomalies in the orbit of Mercury, but it was never found (Baum and Sheehan, 1997). On my view, recent practice in psycholinguistics, I contend, looks more like the search for Vulcan than the search for Neptune. For instance, Yoshida et al. (2026) demonstrate that surprisal theory’s failure to predict garden-path effects (Huang et al., 2024) can be absorbed by fine-tuning the language model on the reading times themselves. Indeed, re-estimating the language model in light of new data gives surprisal theory a clean mechanism to accommodate anomalies. By Proposition 1, a suitable language model always exists. Thus, re-estimation itself is a symptom of the problem, rather than a solution.

7 Conclusion

This paper has argued that surprisal theory, as currently practiced, is tautological. I show that, for any pattern of processing difficulty, a language model can be constructed whose surprisal reproduces it (Proposition 1), so the theory makes no falsifiable prediction until the language model is constrained from outside. Further, in an archaeology of the literature, I identify the corpus assumption, which has implicitly undergirded research in surprisal theory: that the relevant language model is the very distribution that generated the training corpus, so that better corpus fit should yield better predictions of processing difficulty (Proposition 2). This scaling implication is what made surprisal theory falsifiable in practice, and, indeed, it is what recent evidence has falsified. As Oh and Schuler (2023) document, the fit between surprisal and reading times improves as language models grow larger only up to a point after which it falls off—even as the models become better approximations to p_C . All in all, I contend this leaves surprisal theory—at present—without clean grounding, i.e., the language model under which surprisal is computed can neither be left as a free parameter nor be pinned down by any current principle in the literature.

Limitations

This paper is a conceptual and mathematical argument about the logical status of surprisal theory;

it does not present new experiments or computational results. Several limitations follow from this scope. First, the argument that the corpus assumption has been falsified rests on the findings of Oh and Schuler (2023), replicated by Oh et al. (2024) across multiple model families and corpora, and extended to neural measures by Lin and Schuler (2025). While I treat these findings as robust, future work may refine or qualify them—for instance, by disentangling the effects of model architecture from those of corpus fit. Second, the sufficient condition in Proposition 1 is mild and likely satisfied by any realistic difficulty measure, but I have not verified this empirically for specific psychometric datasets. Third, the paper identifies the grounding problem but does not solve it: I sketch two partial approaches (lossy-context surprisal and developmentally plausible corpora) without providing a complete, independently justified specification of q . Fourth, Theorem 1 defines the estimator q_N as the $\arg \max$ of the empirical log-likelihood over $\mathcal{Q}$, and the scaling implication inherits this idealization: the guarantee concerns the model a perfect optimizer *would* find. Perfect optimization is likely unattainable in practice—even training a three-node neural network to fit a dataset is NP-complete (Blum and Rivest, 1992)—and gradient-based training offers no guarantee of reaching the global optimum; bridging this gap would require optimization-aware theory beyond the scope of this paper. Finally, the analogy to the propensity interpretation of fitness in evolutionary biology is meant to be structurally illuminating, not to claim that the two cases are identical in all respects.

Acknowledgements

I thank Eleanor Chodroff, Francesco Ignazio Re, Andreas Opedal, Richard Futrell, Samuel Kiege-land, Taiga Someya, Timothy J. O’Donnell, Tim Vieira, Kate McCurdy, Thomas Hikaru Clark, Jacob Hoover Vigly, and Zach Hopton for helpful discussion and feedback on drafts of this paper.

AI Declaration

Claude Code (Anthropic) was used for proofreading, phrasing, and checking the mathematical arguments. All claims, arguments, and proofs are my own, and I take full responsibility for the content of the paper. Any remaining errors belong to me.

References

- Gerry T. M. Altmann and Yuki Kamide. 1999. **Incremental interpretation at verbs: Restricting the domain of subsequent reference.** *Cognition*, 73(3):247–264.
- John R. Anderson. 1990. *The Adaptive Character of Thought*. Lawrence Erlbaum Associates, Hillsdale, NJ.
- Richard Baum and William Sheehan. 1997. *In Search of Planet Vulcan: The Ghost in Newton’s Clockwork Universe*. Plenum Press, New York.
- Patrick Billingsley. 2012. *Probability and Measure*, anniversary edition. John Wiley & Sons.
- Avrim L. Blum and Ronald L. Rivest. 1992. **Training a 3-node neural network is NP-complete.** *Neural Networks*, 5(1):117–127.
- Taylor L. Booth and Richard A. Thompson. 1973. **Applying probability measures to abstract languages.** *IEEE Transactions on Computers*, C-22(5):442–450.
- Marisa Ferrara Boston, John Hale, Reinhold Kliegl, Umesh Patil, and Shravan Vasishth. 2008. **Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam sentence corpus.** *Journal of Eye Movement Research*, 2(1):1–12.
- Robert N. Brandon. 1978. **Adaptation and evolutionary theory.** *Studies in History and Philosophy of Science*, 9(3):181–206.
- Samuel Butler. 1882. *Evolution, Old and New*, 2nd edition. Hardwicke and Bogue, London.
- Andrés Buxó-Lugo, Aniello De Santo, Morgan Grobol, Ryan J. Hubbard, and Cassandra L. Jacobs. 2026. **surprisal is Not a Theory.** ArXiv:2607.20208.
- Lucas Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Y. Hu, Jing Liu, Jaap Jumelet, Tal Linzen, Aaron Mueller, Candace Ross, Raj Sanjay Shah, Alex Warstadt, Ethan Gotlieb Wilcox, and Adina Williams. 2025. **Findings of the third BabyLM challenge: Accelerating language modeling research with cognitively plausible data.** In *Proceedings of the First BabyLM Workshop*.
- Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA.
- Noam Chomsky. 1966. *Cartesian Linguistics: A Chapter in the History of Rationalist Thought*. Harper & Row, New York.
- Charles Darwin. 1859. *On the Origin of Species by Means of Natural Selection*. John Murray, London.
- Andrea De Varda and Marco Marelli. 2024. **Locally biased transformers better align with human reading times.** In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 30–36.
- Gary S. Dell. 1986. **A spreading-activation theory of retrieval in sentence production.** *Psychological Review*, 93(3):283–321.
- Vera Demberg and Frank Keller. 2008. **Data from eye-tracking corpora as evidence for theories of syntactic processing complexity.** *Cognition*, 109(2):193–210.
- Li Du, Lucas Torroba Hennigen, Tiago Pimentel, Clara Meister, Jason Eisner, and Ryan Cotterell. 2023. **A measure-theoretic characterization of tight language models.** In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 9744–9770.
- Stefan L. Frank, Leun J. Otten, Giulia Galli, and Gabriella Vigliocco. 2015. **The ERP response to the amount of information conveyed by words in sentences.** *Brain and Language*, 140:1–11.
- Richard Futrell, Edward Gibson, and Roger P. Levy. 2020. **Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing.** *Cognitive Science*, 44(3):e12814.
- Mario Giulianelli, Andreas Opedal, and Ryan Cotterell. 2024. **Generalized measures of anticipation and responsivity in online language processing.** In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11648–11669.
- Adam Goodkind and Klinton Bicknell. 2018. **Predictive power of word surprisal for reading times is a linear function of language model quality.** In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics*, pages 10–18.
- Morton Grosser. 1962. *The Discovery of Neptune*. Harvard University Press, Cambridge, MA.
- Michael Hahn, Richard Futrell, Roger P. Levy, and Edward Gibson. 2022. **A resource-rational model of human processing of recursive linguistic structure.** *Proceedings of the National Academy of Sciences*, 119(43):e2122602119.
- John Hale. 2001. **A probabilistic Earley parser as a psycholinguistic model.** *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 159–166.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. **Long short-term memory.** *Neural Computation*, 9(8):1735–1780.
- Michael Y. Hu, Aaron Mueller, Candace Ross, Adina Williams, Tal Linzen, Chengxu Zhuang, Ryan Cotterell, Leshem Choshen, Alex Warstadt, and Ethan Gotlieb Wilcox. 2024. **Findings of the second BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora.** In *Proceedings of the 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 1–21.

- Kuan-Jung Huang, Suhas Arehalli, Mari Kugemoto, Christian Muxica, Grusha Prasad, Brian Dillon, and Tal Linzen. 2024. [Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty](#). *Journal of Memory and Language*, 137:104510.
- Samuel Kiegeand, Vésteinn Snæbjarnarson, Tim Vieira, and Ryan Cotterell. 2026. [On the proper treatment of units in surprisal theory](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32202–32224.
- Samuel Kiegeand, Ethan Wilcox, Afra Amini, David Robert Reich, and Ryan Cotterell. 2024. [Reverse-engineering the reader](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9367–9389.
- Gina R. Kuperberg and T. Florian Jaeger. 2016. [What do we mean by prediction in language comprehension?](#) *Language, Cognition and Neuroscience*, 31(1):32–59.
- Tatsuki Kuribayashi, Yohei Oseki, Souhaib Ben Taieb, Kentaro Inui, and Timothy Baldwin. 2025. [Large language models are human-like internally](#). *Transactions of the Association for Computational Linguistics*, 13:1743–1766.
- Imre Lakatos. 1970. [Falsification and the methodology of scientific research programmes](#). In Imre Lakatos and Alan Musgrave, editors, *Criticism and the Growth of Knowledge*, pages 91–196. Cambridge University Press.
- Roger Levy. 2008. [Expectation-based syntactic comprehension](#). *Cognition*, 106(3):1126–1177.
- Yi-Chien Lin and William Schuler. 2025. [Surprisal from larger transformer-based language models predicts fMRI data more poorly](#). *arXiv preprint*.
- Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*, 3rd edition. Lawrence Erlbaum Associates, Mahwah, NJ.
- Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. [Building a large annotated corpus of English: The Penn Treebank](#). *Computational Linguistics*, 19(2):313–330.
- David Marr. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman and Company, San Francisco, CA.
- Kate McCurdy and Michael Hahn. 2024. [Lossy context surprisal predicts task-dependent patterns in relative clause processing](#). In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 36–45.
- Clara Meister, Tiago Pimentel, Thomas Hikaru Clark, Ryan Cotterell, and Roger P. Levy. 2022. [Analyzing wrap-up effects through an information-theoretic lens](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.
- Susan K. Mills and John H. Beatty. 1979. [The propensity interpretation of fitness](#). *Philosophy of Science*, 46(2):263–286.
- Byung-Doh Oh and William Schuler. 2023. [Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times?](#) *Transactions of the Association for Computational Linguistics*, 11:336–350.
- Byung-Doh Oh, Shisen Yue, and William Schuler. 2024. [Frequency explains the inverse correlation of large language models’ size, training data amount, and surprisal’s fit to reading times](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2644–2663.
- Byung-Doh Oh, Hongao Zhu, and William Schuler. 2025. [The inverse scaling effect of pre-trained language model surprisal is not due to data leakage](#). In *Findings of the Association for Computational Linguistics: ACL 2025*.
- Andreas Opedal, Eleanor Chodroff, Ryan Cotterell, and Ethan Gotlieb Wilcox. 2024. [On the role of context in reading time prediction](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.
- Robert Henry Peters. 1976. [Tautology in evolution and ecology](#). *The American Naturalist*, 110(971):1–12.
- Martin J. Pickering and Simon Garrod. 2004. [Toward a mechanistic psychology of dialogue](#). *Behavioral and Brain Sciences*, 27(2):169–190.
- Karl R. Popper. 1959. *The Logic of Scientific Discovery*. Hutchinson, London.
- Karl R. Popper. 1974. [Darwinism as a metaphysical research programme](#). In Paul Arthur Schilpp, editor, *The Philosophy of Karl Popper*, pages 133–143. Open Court, La Salle, IL.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language models are unsupervised multitask learners](#). Technical report, OpenAI.
- Henry Scheffé. 1947. [A useful convergence theorem for probability distributions](#). *The Annals of Mathematical Statistics*, 18(3):434–438.
- Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger P. Levy. 2024. [Large-scale evidence for logarithmic effects of word predictability on reading time](#). *Proceedings of the National Academy of Sciences*, 121(10):e2307876121.

- Edmund W. Sinnott, L. C. Dunn, and Theodosius Dobzhansky. 1958. *Principles of Genetics*, 5th edition. McGraw-Hill, New York.
- Nathaniel J. Smith and Roger Levy. 2013. *The effect of word predictability on reading time is logarithmic*. *Cognition*, 128(3):302–319.
- Vésteinn Snæbjarnarson, Samuel Kiegeland, Tianyu Liu, Reda Boumasmoud, Ryan Cotterell, and Tim Vieira. 2026. *Transducing language models*. In *Proceedings of the International Conference on Learning Representations*.
- Michael K. Tanenhaus, Michael J. Spivey-Knowlton, Kathleen M. Eberhard, and Julie C. Sedivy. 1995. *Integration of visual and linguistic information in spoken language comprehension*. *Science*, 268(5217):1632–1634.
- William Timkey and Tal Linzen. 2023. *A language model with limited memory capacity captures interference in human sentence processing*. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.
- Eleftheria Tsipidi, Samuel Kiegeland, Francesco Ignazio Re, Tianyang Xu, Mario Giulianelli, Karolina Stańczak, and Ryan Cotterell. 2026. *Probing for reading times*. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12618–12642.
- Aad W. van der Vaart. 1998. *Asymptotic Statistics*. Cambridge University Press.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. *Attention is all you need*. In *Advances in Neural Information Processing Systems*, volume 30.
- Conrad Hal Waddington. 1960. *Evolutionary adaptation*. In Sol Tax, editor, *Evolution After Darwin*, volume 1, pages 381–402. University of Chicago Press, Chicago.
- Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Gotlieb Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. *Findings of the BabyLM challenge: Sample-efficient pre-training on developmentally plausible corpora*. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34.
- Halbert White. 1982. *Maximum likelihood estimation of misspecified models*. *Econometrica*, 50(1):1–25.
- Halbert White. 1989. *Learning in artificial neural networks: A statistical perspective*. *Neural Computation*, 1(4):425–464.
- Ethan Gotlieb Wilcox, Tiago Pimentel, Clara Meister, and Ryan Cotterell. 2024. *An information-theoretic analysis of targeted regressions during reading*. *Cognition*, 249:105765.
- Ethan Gotlieb Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. *Testing the predictions of surprisal theory in 11 languages*. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.
- Ryo Yoshida, Shinnosuke Isono, Taiga Someya, Yohei Oseki, and Tatsuki Kuribayashi. 2026. *An existence proof for neural language models that can explain garden-path effects via surprisal*. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36556–36569.

Contents of the Appendix

A	Consistency of the MLE for Language Models	14
B	Deferred Proofs	16
B.1	Proof of Proposition 1	16
B.2	Proof of Lemma 1	16
B.3	Proof of Proposition 2	17

A Consistency of the MLE for Language Models

This appendix gives a self-contained proof that the maximum-likelihood estimator converges to the KL projection of the data-generating distribution onto the model family. A proof is included here, rather than merely cited, because the standard references (White, 1982, 1989; van der Vaart, 1998) do not explicitly verify the uniform domination condition for language models, i.e., distributions over Σ^* . As I show below, this condition reduces to the assumption that p_C has finite expected length. The result below is stated entirely in function space, i.e., the family $\mathcal{Q}$ is treated as a compact set of language models, with no reference to an underlying parameter space. Following White (1989), the theorem shows convergence to the *set* of KL-minimizers, so no identification assumption is required, which is essential for neural language models, whose parameter symmetries (e.g., permutation of attention heads) make point identification impossible.

An important caveat: the theorem assumes that the MLE q_N is a *global* maximizer of the empirical log-likelihood over $\mathcal{Q}$. In practice, neural language models are trained by stochastic gradient descent on a non-convex loss landscape, and the optimizer may converge to a local rather than global optimum (see footnote 18 in the main text). Additionally, they are often not trained to converged at all. The scaling implication, and hence the consistency result, is therefore a statement about the population-level optimum within $\mathcal{Q}$ —the best model that *could* be found—not about the output of any particular training run.

Theorem 1 (Consistency of the MLE). *Let Σ be an alphabet with $|\Sigma| = V$, and let p_C be a data-generating distribution over Σ^* . Next, let $\mathcal{Q}$ be a family of language models over Σ^* , equipped with the topology of pointwise convergence.²³ Additionally, let $\mathcal{Q}^* \stackrel{\text{def}}{=} \arg \min_{q \in \mathcal{Q}} D_{\text{KL}}(p_C \parallel q)$ be the set of KL-minimizers, let $Q(q) \stackrel{\text{def}}{=} \mathbb{E}_{\mathbf{u} \sim p_C} [\log(q(\mathbf{u}))]$ denote the expected log-likelihood, and let q_N denote the maximum-likelihood estimator of p_C within $\mathcal{Q}$ computed from an i.i.d. corpus $\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(N)} \sim p_C$. If the following three conditions hold:*

- (i) **Compactness.** $\mathcal{Q}$ is compact.
- (ii) **Uniform domination.** There exists a p_C -integrable function $d: \Sigma^* \rightarrow \mathbb{R}_{\geq 0}$ such that $\sup_{q \in \mathcal{Q}} |\log(q(\mathbf{u}))| \leq d(\mathbf{u})$ for all $\mathbf{u} \in \Sigma^*$.
- (iii) **Finite expected length.** $\mathbb{E}_{\mathbf{u} \sim p_C} [|\mathbf{u}|] < \infty$.

Then, almost surely, every limit point of the MLE sequence $\{q_N\}_{N=1}^\infty$ lies in $\mathcal{Q}^*$. In particular,

$$D_{\text{KL}}(p_C \parallel q_N) \xrightarrow[N \rightarrow \infty]{a.s.} \inf_{q \in \mathcal{Q}} D_{\text{KL}}(p_C \parallel q). \quad (13)$$

If $\mathcal{Q}^*$ is a singleton $\{q^*\}$, then $q_N \rightarrow q^*$ pointwise a.s.

Discussion of the assumptions. For parametric families—such as neural language models with bounded weights—compactness of the parameter space and continuity of the parameterization give compactness of $\mathcal{Q}$ in the pointwise topology, provided all models in $\mathcal{Q}$ are tight. Showing tightness, however, is non-trivial: even with a compact parameter space, a sequence of language models can converge pointwise to a non-tight model if the expected length diverges along the sequence. Because $[0, 1]^{\Sigma^*}$ is compact in the pointwise topology by Tychonoff’s theorem, $\mathcal{Q}$ is compact if and only if it is closed in $[0, 1]^{\Sigma^*}$, i.e., if and only if every pointwise limit of members of $\mathcal{Q}$ is again a member of $\mathcal{Q}$ and, in particular, a language model. Note that finiteness of each model’s expected length does not suffice—expected lengths may diverge along a sequence. A *uniform* bound, by contrast, does suffice. Suppose $\mathbb{E}_{\mathbf{u} \sim q} [|\mathbf{u}|] \leq L$ for every $q \in \mathcal{Q}$, and let $q_n \rightarrow q_\infty$ pointwise, where a priori q_∞ is only a sub-probability measure on Σ^* .²⁴

²³Formally, $q_n \rightarrow q$ in $\mathcal{Q}$ if and only if $q_n(\mathbf{u}) \rightarrow q(\mathbf{u})$ for every $\mathbf{u} \in \Sigma^*$. Because Σ^* is countable, the topology of pointwise convergence is metrizable, so sequential continuity and continuity coincide on $\mathcal{Q}$.

²⁴That q_∞ has total mass at most 1 follows from Fatou’s lemma applied to counting measure on Σ^* . In particular, $\sum_{\mathbf{u} \in \Sigma^*} q_\infty(\mathbf{u}) = \sum_{\mathbf{u} \in \Sigma^*} \lim_{n \rightarrow \infty} q_n(\mathbf{u}) \leq \liminf_{n \rightarrow \infty} \sum_{\mathbf{u} \in \Sigma^*} q_n(\mathbf{u}) = 1$.

For any $T \in \mathbb{Z}_{>0}$, only finitely many $\mathbf{u} \in \Sigma^*$ satisfy $|\mathbf{u}| \leq T$, so pointwise convergence gives

$$\sum_{\mathbf{u} \in \Sigma^*} \mathbb{1}\{|\mathbf{u}| \leq T\} q_\infty(\mathbf{u}) = \lim_{n \rightarrow \infty} \sum_{\mathbf{u} \in \Sigma^*} \mathbb{1}\{|\mathbf{u}| \leq T\} q_n(\mathbf{u}) = \lim_{n \rightarrow \infty} \left(1 - \Pr_{\mathbf{u} \sim q_n}(|\mathbf{u}| > T)\right) \geq 1 - \frac{L}{T}, \quad (14)$$

where the inequality is Markov's. Letting $T \rightarrow \infty$ yields $\sum_{\mathbf{u} \in \Sigma^*} q_\infty(\mathbf{u}) = 1$, i.e., q_∞ is a language model. For (ii), if every $q \in \mathcal{Q}$ assigns each conditional probability at least $\eta > 0$ (the standing assumption of footnote 16 in Definition 1, which softmax-based models on a compact parameter space satisfy), then $|\log(q(\mathbf{u}))| \leq (|\mathbf{u}| + 1) \log(1/\eta)$, and the domination condition reduces to (iii). The same assumption yields the uniform expected-length bound $\mathbb{E}_{\mathbf{u} \sim q}[|\mathbf{u}|] \leq (1 - \eta)/\eta$ across $\mathcal{Q}$ (by the geometric argument in the proof of Proposition 1), which, by the argument just given, ensures that pointwise limits of members of $\mathcal{Q}$ are language models. The finite expected length assumption (iii) plays a dual role: it ensures the domination condition (ii) for the consistency theorem and the well-definedness of the fitness measure M (Definition 1), since both require that sums over prefixes weighted by $\vec{p}_C$ converge. This is the key assumption not accounted for by the standard references (White, 1982; van der Vaart, 1998), which work in finite-dimensional parameter spaces and do not consider the structure of distributions over Σ^* .

Proof. Define the empirical log-likelihood $Q_N(q) \stackrel{\text{def}}{=} \frac{1}{N} \sum_{n=1}^N \log(q(\mathbf{u}^{(n)}))$, so that $q_N = \arg \max_{q \in \mathcal{Q}} Q_N(q)$. Note that $Q(q) = -D_{\text{KL}}(p_C \| q) - H(p_C)$, where $H(p_C)$ is finite because finite expected length (assumption (iii)) implies $H(p_C) \leq (\mathbb{E}_{\mathbf{u} \sim p_C}[|\mathbf{u}|] + 1) \log |\Sigma_E| < \infty$; hence maximizing Q is equivalent to minimizing KL divergence. The argument follows White (1982, 1989) and proceeds in four steps.

- **Step 1: The expected log-likelihood Q is finite.** For each $q \in \mathcal{Q}$, the integrand $|\log(q(\mathbf{u}))|$ is bounded by the dominating function $d(\mathbf{u})$ from assumption (ii). Since d is p_C -integrable by assumption, $\mathbb{E}_{\mathbf{u} \sim p_C}[|\log(q(\mathbf{u}))|] \leq \mathbb{E}_{\mathbf{u} \sim p_C}[d(\mathbf{u})] < \infty$. Therefore $Q(q) = \mathbb{E}_{\mathbf{u} \sim p_C}[\log(q(\mathbf{u}))]$ is finite for every $q \in \mathcal{Q}$.
- **Step 2: Q is continuous on $\mathcal{Q}$.** Let $\{q_n\}_{n=1}^\infty$ be a sequence in $\mathcal{Q}$ with $q_n \rightarrow q$ pointwise. Then for each $\mathbf{u} \in \Sigma^*$, $\log(q_n(\mathbf{u})) \rightarrow \log(q(\mathbf{u}))$ by continuity of the logarithm. By assumption (ii), $|\log(q_n(\mathbf{u}))| \leq d(\mathbf{u})$ for all n and all $\mathbf{u}$, where d is p_C -integrable. The dominated convergence theorem (Billingsley, 2012) then gives $Q(q_n) = \mathbb{E}_{\mathbf{u} \sim p_C}[\log(q_n(\mathbf{u}))] \rightarrow \mathbb{E}_{\mathbf{u} \sim p_C}[\log(q(\mathbf{u}))] = Q(q)$, so Q is sequentially continuous, and hence continuous, on $\mathcal{Q}$, because the topology of pointwise convergence on $\mathcal{Q}$ is metrizable (footnote 23).
- **Step 3: Uniform law of large numbers.** The strong law of large numbers gives $Q_N(q) \rightarrow Q(q)$ a.s. for each fixed $q \in \mathcal{Q}$. I upgrade this to *uniform* convergence over $\mathcal{Q}$. The family of functions $\{\mathbf{u} \mapsto \log(q(\mathbf{u})) : q \in \mathcal{Q}\}$ is dominated by the p_C -integrable function d (assumption (ii)). By assumption (i), $\mathcal{Q}$ is compact, and it is metrizable (footnote 23); moreover, the map $q \mapsto \log(q(\mathbf{u}))$ is continuous for each fixed $\mathbf{u}$, as established in the proof of Step 2. The family is therefore a p_C -Glivenko–Cantelli class (van der Vaart, 1998, Example 19.8), which yields

$$\sup_{q \in \mathcal{Q}} |Q_N(q) - Q(q)| \xrightarrow[N \rightarrow \infty]{\text{a.s.}} 0. \quad (15)$$

- **Step 4: Convergence of maximizers.** The maximizer q_N exists because Q_N is a finite sum of continuous functions on the compact set $\mathcal{Q}$. The uniform convergence of Step 3 holds on an event of probability one. Fix a realization of this event $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \dots$. Then, let $q^\dagger$ be any limit point of the sequence $\{q_N\}_{N=1}^\infty$. By compactness (i), at least one such limit point exists. Passing to a subsequence $\{q_{N_k}\}$ with $q_{N_k} \rightarrow q^\dagger$, the uniform convergence from Step 3 gives

$$|Q_{N_k}(q_{N_k}) - Q(q^\dagger)| \leq \underbrace{|Q_{N_k}(q_{N_k}) - Q(q_{N_k})|}_{\rightarrow 0 \text{ by Step 3}} + \underbrace{|Q(q_{N_k}) - Q(q^\dagger)|}_{\rightarrow 0 \text{ by Step 2}} \rightarrow 0. \quad (16)$$

Now, since q_{N_k} maximizes Q_{N_k} over $\mathcal{Q}$, for any $q \in \mathcal{Q}$, it holds that $Q_{N_k}(q_{N_k}) \geq Q_{N_k}(q)$. Taking $k \rightarrow \infty$ and applying the uniform convergence of Step 3 to both sides, $Q(q^\dagger) \geq Q(q)$ for every $q \in \mathcal{Q}$. Hence $q^\dagger \in \arg \max_{q \in \mathcal{Q}} Q(q) = \mathcal{Q}^*$. Since $Q(q) = -D_{\text{KL}}(p_C \parallel q) + \text{const}$, this is equivalent to $q^\dagger \in \arg \min_{q \in \mathcal{Q}} D_{\text{KL}}(p_C \parallel q)$. Because every limit point of $\{q_N\}$ lies in $\mathcal{Q}^*$, compactness of $\mathcal{Q}$ and continuity of Q (Step 2) imply that the KL divergence converges: $D_{\text{KL}}(p_C \parallel q_N) \rightarrow \inf_{q \in \mathcal{Q}} D_{\text{KL}}(p_C \parallel q)$ a.s. If $\mathcal{Q}^* = \{q^*\}$ is a singleton, then every limit point equals q^* , and by compactness the full sequence converges: $q_N \rightarrow q^*$ pointwise a.s. ■

B Deferred Proofs

B.1 Proof of Proposition 1

Proposition 1. Let $d: \Sigma_E \times \Sigma^* \rightarrow \mathbb{R}_{\geq 0}$ be a non-negative difficulty measure, and let q_d be defined by Eq. (6). If there exists a function $f: \mathbb{Z}_{>0} \rightarrow \mathbb{R}_{\geq 0}$ such that

$$d(\text{EOS} \mid \mathbf{u}) \leq f(|\mathbf{u}| + 1) \quad \text{for all } \mathbf{u} \in \Sigma^* \quad (7)$$

and $\sum_{t=1}^{\infty} \exp(-f(t)) = \infty$, then (a) q_d is tight (see §2), i.e., $\sum_{\mathbf{u} \in \Sigma^*} q_d(\mathbf{u}) = 1$; and (b) if moreover f is constant—i.e., there exists $C > 0$ such that $d(\text{EOS} \mid \mathbf{u}) \leq C$ for all $\mathbf{u} \in \Sigma^*$ —then q_d has finite expected length: $\mathbb{E}_{\mathbf{u} \sim q_d}[|\mathbf{u}|] < \infty$.

Proof. Part (a). Each factor in Eq. (6) defines a valid conditional distribution over Σ_E (by the softmax normalization), so q_d is a conditional collection in the sense of §2. I apply Du et al. (2023, Proposition 4.3), which states that a conditional collection is tight if there exists a function $h: \mathbb{Z}_{>0} \rightarrow \mathbb{R}_{\geq 0}$ satisfying (i) $\vec{q}_d(\text{EOS} \mid \mathbf{u}) \geq h(|\mathbf{u}| + 1)$ for all $\mathbf{u} \in \Sigma^*$, and (ii) $\sum_{t=1}^{\infty} h(t) = \infty$. I construct such an h . Since $d \geq 0$, each term in the denominator of Eq. (6) satisfies $\exp(-d(u' \mid \mathbf{u})) \leq 1$, so the denominator is at most $|\Sigma_E|$. The numerator for EOS is $\exp(-d(\text{EOS} \mid \mathbf{u}))$. Therefore:

$$\vec{q}_d(\text{EOS} \mid \mathbf{u}) = \frac{\exp(-d(\text{EOS} \mid \mathbf{u}))}{\sum_{u' \in \Sigma_E} \exp(-d(u' \mid \mathbf{u}))} \quad (17a)$$

$$\geq \frac{\exp(-d(\text{EOS} \mid \mathbf{u}))}{|\Sigma_E|}. \quad (17b)$$

By hypothesis, $d(\text{EOS} \mid \mathbf{u}) \leq f(|\mathbf{u}| + 1)$, so $\exp(-d(\text{EOS} \mid \mathbf{u})) \geq \exp(-f(|\mathbf{u}| + 1))$. Setting $h(t) \stackrel{\text{def}}{=} \exp(-f(t))/|\Sigma_E|$, condition (i) is satisfied. For condition (ii): $\sum_{t=1}^{\infty} h(t) = \frac{1}{|\Sigma_E|} \sum_{t=1}^{\infty} \exp(-f(t)) = \infty$ by hypothesis. Hence, q_d is tight.

Part (b). If f is the constant function $f(t) = C$, then $\sum_{t=1}^{\infty} \exp(-f(t)) = \infty$, so q_d is tight by part (a). Moreover, Eq. (17b) gives $\vec{q}_d(\text{EOS} \mid \mathbf{u}) \geq \exp(-C)/|\Sigma_E|$ for all $\mathbf{u}$. Setting $\rho \stackrel{\text{def}}{=} 1 - \exp(-C)/|\Sigma_E| < 1$, the probability that a string sampled from q_d has length at least t is at most ρ^t , since such a string must fail to generate EOS at each of its first t positions. Therefore, $\mathbb{E}_{\mathbf{u} \sim q_d}[|\mathbf{u}|] = \sum_{t=1}^{\infty} \Pr(|\mathbf{u}| \geq t) \leq \sum_{t=1}^{\infty} \rho^t = \rho/(1 - \rho) < \infty$. ■

B.2 Proof of Lemma 1

Lemma 1 (Continuity of M). Suppose p has finite expected length and g is continuous and bounded. Then $M: \mathcal{Q} \rightarrow \mathbb{R}$ is continuous.

Proof. Suppose $q_n \rightarrow q$ pointwise, i.e., $q_n(\mathbf{u}) \rightarrow q(\mathbf{u})$ for all $\mathbf{u} \in \Sigma^*$. Because every member of $\mathcal{Q}$ is a language model—a probability distribution over Σ^* with total mass 1—Scheffé’s lemma (Scheffé, 1947) upgrades pointwise convergence to convergence in L^1 , i.e., $\sum_{\mathbf{u} \in \Sigma^*} |q_n(\mathbf{u}) - q(\mathbf{u})| \rightarrow 0$. Each prefix probability is a sum of string probabilities, so $\vec{q}_n(\mathbf{u}) \rightarrow \vec{q}(\mathbf{u})$ for every $\mathbf{u} \in \Sigma^*$; and, because per-unit probabilities are bounded away from zero (Definition 1), the limiting denominators satisfy $\vec{q}(\mathbf{u}) > 0$. Hence the conditionals converge: $\vec{q}_n(u \mid \mathbf{u}) \rightarrow \vec{q}(u \mid \mathbf{u})$ for all $u \in \Sigma_E$ and $\mathbf{u} \in \Sigma^*$. Because g is continuous, the integrand converges pointwise for each fixed $u \in \Sigma_E$ and $\mathbf{u} \in \Sigma^*$. Let B be a

bound on $|g|$, which exists by assumption. Because the conditionals sum to one, $\sum_{u \in \Sigma_E} \vec{p}(uu) = \vec{p}(u)$, and, because p has finite expected length, $\sum_{u \in \Sigma^*} \vec{p}(u) = \mathbb{E}_{u \sim p}[|u|] + 1 < \infty$ (Opedal et al., 2024, Proposition 1); hence the weights $\vec{p}(uu)$ are summable and $B \cdot \vec{p}(uu)$ is an integrable dominating function. Thus, for each u and u , the integrand $\vec{p}(uu) g(d(u | u), -\log(\vec{q}_n(u | u)))$ converges to $\vec{p}(uu) g(d(u | u), -\log(\vec{q}(u | u)))$, and all terms are bounded in absolute value by $B \cdot \vec{p}(uu)$, which is summable. The dominated convergence theorem (Billingsley, 2012) then permits interchange of the limit and the sum, yielding $M(q_n) \rightarrow M(q)$. Because the pointwise topology on $\mathcal{Q}$ is metrizable (footnote 23 in App. A), sequential continuity is equivalent to continuity, which completes the proof. ■

B.3 Proof of Proposition 2

Proposition 2 (Scaling implication). *Assume $\mathcal{Q}^*$ is a singleton $\{q^*\}$, i.e., the KL-minimizer is unique. Under the regularity conditions of Theorem 1, it holds that*

$$M(q_N) \xrightarrow[N \rightarrow \infty]{p} M(q^*). \quad (11)$$

In particular, for any $\epsilon > 0$ and $\delta > 0$, there exists N_0 such that for all $N \geq N_0$, it holds that $\Pr(|M(q_N) - M(q^*)| < \epsilon) > 1 - \delta$.

Proof. Because $\mathcal{Q}^*$ is a singleton, Theorem 1 gives $q_N \xrightarrow[N \rightarrow \infty]{\text{a.s.}} q^*$ pointwise. By Lemma 1, applied with $p = p_C$, which has finite expected length by assumption (iii) of Theorem 1, M is continuous on $\mathcal{Q}$. Hence, $M(q_N) \xrightarrow[N \rightarrow \infty]{\text{a.s.}} M(q^*)$ by the continuous mapping theorem (cf. van der Vaart, 1998, Theorem 2.3), and almost-sure convergence implies convergence in probability. ■