Estimating Surprisal from the Speech Signal

Zachary Hopton[1] **Samuel Kiegeland**[2,3] **Francesco Ignazio Re**[2] **Patrick Haller**[1]
Matthew B. Winn[4] **Ryan Cotterell**[2] **Eleanor Chodroff**[1]

[1]University of Zurich [2]ETH Zürich [3]CHI-FRO [4]University of Minnesota
{zacharywilliam.hopton, eleanor.chodroff}@uzh.ch

Abstract

Although language is predominantly spoken, empirical tests of surprisal theory rely almost entirely on written text. In practice, surprisal is estimated using language models (LMs) trained on text and applied to transcriptions, creating a text bottleneck that limits evaluation of the theory in the spoken modality. We remove this bottleneck by estimating surprisal directly from the speech signal using a speech LM. This raises a key question: should surprisal be defined over realization-specific acoustic inputs or over abstract linguistic units? Since a single unit can have many distinct acoustic realizations, each assigned a different probability by the model, defining surprisal over abstract units is difficult. To address this, we introduce an importance sampling estimator that approximates the probability of an abstract unit by marginalizing over realizations generated by a text-to-speech model. Using pupillometry data from a listening study, we evaluate surprisal theory in this text-free setting at multiple levels of granularity. Obtaining surprisal from several speech LMs, we consistently find that surprisal estimated via importance sampling better predicts pupil dilation than realization-specific surprisal for words. We also find that surprisal values for sub-lexical units are predictive of pupil dilation. While we find that surprisal can be estimated without recourse to text, and that marginalizing over distinct realizations of words is more predictive of lexical processing effort, surprisal as a whole is not a strong predictor of change in pupil size, regardless of its source.

hopton-etal-2026-speech

1 Introduction

Human language is primarily a spoken phenomenon, yet empirical tests of surprisal theory rely on written text rather than the acoustic signal itself (Hale, 2001; Levy, 2008; Shain et al., 2024).[1] In studies of reading, this assumption is natural: the input is discretized into written units, and text language models (LMs) can assign probabilities to those units directly. In studies of listening, however, the input is an acoustic signal. Applying a text language model to a transcription of the acoustic signal introduces a text bottleneck: the probabilistic model is defined over written forms, while listeners receive and interpret speech. This bottleneck is not only an inconvenience; it also obscures a fundamental question for calculating surprisal for a speech signal from a speech LM: what is the unit over which surprisal should be calculated?

Existing work on spoken-language surprisal has generally followed the text-based approach by applying text LMs to transcriptions (Shain et al., 2020; Wallbridge et al., 2022). This assumes that textual representations are adequate proxies for the units used in listening. Even when these transcriptions are in terms of units that are more representative of speech than graphemes (e.g., phonemes or syllables as in Pimentel et al. (2021) and Sohoglu et al. (2024)), transcription omits prosodic, acoustic, and speaker-specific information present in the speech signal. Speech LMs remove this bottleneck, but introduce a new problem: their probabilities are defined over speech-code sequences rather than words or phones.

We address this problem using pupillometry data from Winn (2016), where participants listened to sentences varying in semantic context. As these stimuli have minimal phonetic variation (one speaker, flat affect, read speech prosody), they provide a controlled auditory test case for asking which speech-LM surprisal quantity best predicts processing effort. Studying surprisal theory through the lens of speech then raises two major questions:

[1]We note that signed languages constitute another major modality of human language production and perception (Sandler and Lillo-Martin, 2006). The challenges we discuss regarding the mismatch between text-based models and non-textual modalities apply equally to sign.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 54489–54509
October 24-29, 2026 ©2026 Association for Computational Linguistics

Abstract units or specific acoustic realizations? Surprisal theory is typically defined over abstract units such as words or phonemes (Levy, 2008; Smith and Levy, 2013), whereas speech LMs assign probabilities to acoustic representations.[2] This discrepancy between the abstract and the specific is known in speech as the lack of invariance problem: a single abstract linguistic unit can have infinitely many acoustic realizations, and the same unit may be realized differently depending on its phonetic context (Liberman et al., 1967), speaker characteristics (Peterson and Barney, 1952), recording quality, background noise, and conditions of the physical environment (e.g., reverberation, echo, etc.), among other sources of variation. As a result, the speech LM assigns different probabilities to different realizations of the same abstract unit. Aggregating probabilities over the observed utterance therefore yields surprisal for that particular realization, not necessarily for the abstract unit. In the present work, we approximate abstract-unit surprisal by marginalizing over possible realizations with importance sampling (Robert and Casella, 2004).

Which granularity of speech? Speech is continuous, but speech LMs operate over discrete frames. Surprisal can therefore be computed at the frame level, or aggregated over more conventional units such as phones or orthographic words. We calculate surprisal at each of these granularities to ask which levels of speech LM predictions can model listening effort.

In our empirical analysis, we investigate the relationship between pupil dilation and speech-LM surprisal estimated at the word, phone, and frame levels using four speech LMs. At the word level, we further compare realization-specific surprisal against our marginalized, abstract-unit estimate. We find that marginalized word surprisal better predicts pupil dilation than realization-specific word surprisal, though phone- and frame-level surprisal are also predictive. Though surprisal is a weak predictor of pupil response as a whole, our results suggest that in spoken language comprehension, processing effort is better captured by probabilities over abstract lexical units than by probabilities over specific acoustic realizations alone. Our main contribution, however, is methodological: we provide a framework for defining the relevant probabilistic event when obtaining surprisal from speech LMs. Applying this framework to diverse speech conditions stands to position generative speech LMs as tools for learning how factors such as speech content, speaker characteristics, and prosody interact to influence speech processing.

[2]We use the term **abstract units** to refer to any given unit that might compose one's intended message when communicating, as opposed to the units that make up noisy observations of said message. See Vigly et al. (2025) for relevant discussion of noisy-channel inference models of language comprehension.

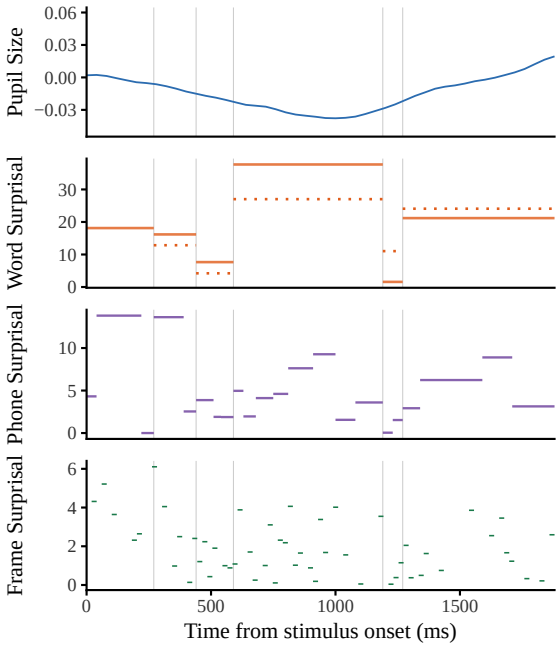

Figure 1: Pupil size trajectory and GSLM unit surprisals for participant NH–24 while listening to stimulus 5x39, "Tom has been discussing the beads." Dotted lines represent estimated marginalized surprisals; solid lines represent unmarginalized surprisals for this specific utterance.

2 Language Models

2.1 Preliminaries

We denote an alphabet—some finite, non-empty set of units—using uppercase Greek letters, e.g., Γ. We use lowercase Greek letters, e.g., $\gamma \in \Gamma$, to denote units from an alphabet and bold lowercase Greek letters $\boldsymbol{\gamma} = \gamma_1 \cdots \gamma_T$ to refer to a string of length T from Γ^*. We use the notation $|\boldsymbol{\gamma}|$ to refer to the length of a string. The concatenation of two strings $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'$ is denoted as $\boldsymbol{\gamma}\boldsymbol{\gamma}'$. $\boldsymbol{\gamma} \prec \boldsymbol{\gamma}'$ denotes that $\boldsymbol{\gamma}$ is a strict prefix of $\boldsymbol{\gamma}'$, and $\boldsymbol{\gamma} \preceq \boldsymbol{\gamma}'$ denotes that $\boldsymbol{\gamma}$ is a prefix of $\boldsymbol{\gamma}'$.

A language model p is a probability distribution over some alphabet. We denote the specific alphabet over which a language model is defined using subscripts, e.g., p_Γ, and extend the alphabet with a

special end-of-sequence symbol EOS $\notin \Gamma$. We use the random variable X that takes values of strings $\gamma \in \Gamma^*$ according to the distribution specified by the LM. The LM's *prefix probability* is defined as

$$\overrightarrow{p_\Gamma}(\gamma) \stackrel{\text{def}}{=} \sum_{\gamma' \in \Gamma^*} p_\Gamma(\gamma\gamma'), \tag{1}$$

the total probability mass of all sequences that begin with the prefix γ. The conditional prefix probability is then

$$\overrightarrow{p_\Gamma}(\gamma' \mid \gamma) \stackrel{\text{def}}{=} \frac{\overrightarrow{p_\Gamma}(\gamma\gamma')}{\overrightarrow{p_\Gamma}(\gamma)}, \tag{2}$$

which defines a valid probability distribution over $\Gamma \sqcup \{\text{EOS}\}$ for all $\gamma \in \Gamma^*$. This structure yields the standard left-to-right factorization of an autoregressive language model:

$$p_\Gamma(\gamma) = \overrightarrow{p_\Gamma}(\text{EOS} \mid \gamma) \prod_{t=1}^{|\gamma|} \overrightarrow{p_\Gamma}(\gamma_t \mid \gamma_{<t}). \tag{3}$$

2.2 Speech Language Models

Whereas text contains natural boundaries between units (e.g., spaces and punctuation in many languages' scripts), speech is carried by a continuous acoustic signal without natural boundaries. Digitization of the speech signal yields a dense sequence of bit-valued amplitude measurements (e.g., 16,000 samples per second). Although these measurements are discrete in a technical sense, treating them as independent units for language modeling is poorly aligned with linguistic structure, as strong dependencies between adjacent samples reflect low-level signal properties rather than meaningful linguistic variation. In practice, speech LMs therefore operate over higher-level representations obtained by discretizing the signal using feature extractors or neural codecs (Arora et al., 2025). An input acoustic signal is first passed through a feature extractor, yielding a shorter sequence of representations. These representations can then be quantized into discrete units from a finite set Δ. As described by Lakhotia et al. (2021), strings of these units from Δ can be used to train an autoregressive language model in the same way as strings of text tokens.

Specifically, we define δ as a string from Δ^*. By defining a speech LM p_Δ as a probability distribution over Δ^* extended with an end-of-sequence symbol EOS $\notin \Delta$, we obtain a left-to-right factorization of an autoregressive language model analogous to that in Eq. (3):

$$p_\Delta(\delta) = \overrightarrow{p_\Delta}(\text{EOS} \mid \delta) \prod_{t=1}^{|\delta|} \overrightarrow{p_\Delta}(\delta_t \mid \delta_{<t}). \tag{4}$$

2.3 Converting Speech LMs to Text LMs

Text can be straightforwardly modeled by defining Σ as an alphabet of units such as orthographic words. Prior work has studied the question of alphabet transfer in LMs, for example, to obtain a character interface from a token-level model (Vieira et al., 2025; Snæbjarnarson et al., 2026). Similarly, we seek to model sequences of units such as orthographic words from a speech LM defined over a different alphabet, i.e., an alphabet of speech tokens. This requires a function $\Omega\colon \Delta^* \to \Sigma^*$ that decodes speech strings to orthographic words. Intuitively, it could be conceptualized as a human listener or automatic speech recognition (ASR) model transcribing an utterance δ. Specifically, we assume that Ω is a *prefix monotone* mapping—extending the input speech (e.g., adding a new speech token to the input sequence) can only leave the output transcription unchanged or cause it to grow; the transcription of any prior prefix is never revised. Though this is a necessary assumption for our approach to obtaining probabilities of abstract units from a speech LM (see §A), there are limits to the extent it realistically applies to speech processing. For example, given the input utterance "/aɪskrim/", Ω is required to commit to transcribing either "I scream" or "Ice cream", though the two could only be disambiguated with further context.

Using the mapping function Ω, words are generated by the speech LM according to the following distribution:

$$p_\Sigma(\sigma) = \mathbb{P}_{X \sim p_\Delta}[\Omega(X) = \sigma]. \tag{5}$$

The probability of some orthographic string σ is then calculated as

$$p_\Sigma(\sigma) = \sum_{\delta \in \mathcal{P}(\sigma)} \mathbb{1}\{\Omega(\delta) = \sigma\} p_\Delta(\delta), \tag{6}$$

where the precover $\mathcal{P}(\sigma) \stackrel{\text{def}}{=} \{\delta \in \Delta^* \colon \Omega(\delta) \succeq \sigma\}$ is the set of all speech strings whose transcription under Ω has σ as a prefix (Vieira et al., 2025). Calculating Eq. (6) is not feasible given that we cannot readily enumerate all possible δ. This being the case, we derive an estimator to approximate prefix probabilities of our units of interest (orthographic words) from a speech LM in §A.

3 Estimating Unit Surprisal

We aim to investigate the relationship between next-unit surprisal under a speech LM and processing effort, as measured by changes in pupil size. Surprisal is computed for multiple units that vary along two dimensions: level of abstraction (abstract units versus surface realizations) and granularity of the acoustic signal (orthographic words, phones, and frames). This yields four surprisal estimates in total: abstracted word-level surprisal obtained via the estimator introduced in §A, and word-, phone-, and frame-level surprisal computed directly over the acoustic signal (see Fig. 1). Below we discuss the implementation of the estimator used to obtain surprisal over abstract units.

3.1 Importance Sampling

We derive an importance sampling estimator $\widehat{\overrightarrow{p_{\Sigma}}}_{,N}(\sigma_t \mid \boldsymbol{\sigma}_{<t})$ that allows us to approximate the conditional probability of an abstract orthographic word from a speech LM (see §A). In importance sampling, we do not sample directly from a speech language model (e.g., randomly sampling the next token given some context), as this would rarely result in the unit of interest σ_t. Rather, we draw speech strings from a proposal distribution $q_{\triangle}$ and reweight them by the ratio $p_{\triangle}/q_{\triangle}$ (Robert and Casella, 2004). Due to Proposition 1 and linearity of expectation, the unconditional importance sampling estimator that we derive is unbiased, i.e., the expectation of the estimator equals the quantity of interest $\overrightarrow{p_{\Sigma}}(\sigma)$. Given N i.i.d. samples $\boldsymbol{\delta}^{(1)}, \ldots, \boldsymbol{\delta}^{(N)} \sim q_{\triangle}$, we first estimate the prefix probability of a text string $\boldsymbol{\sigma}$ as

$$\widehat{\overrightarrow{p_{\Sigma}}}_{,N}(\boldsymbol{\sigma}) \stackrel{\text{def}}{=} \frac{1}{N}\sum_{n=1}^{N}\left[\mathbb{1}\left\{\boldsymbol{\delta}^{(n)} \in \mathcal{P}(\boldsymbol{\sigma})\right\} \times \frac{\overrightarrow{p_{\triangle}}\left(\pi_{\boldsymbol{\sigma}}\left(\boldsymbol{\delta}^{(n)}\right)\right)}{\overrightarrow{q_{\triangle}}\left(\pi_{\boldsymbol{\sigma}}\left(\boldsymbol{\delta}^{(n)}\right)\right)}\right], \quad (7)$$

where $\pi_{\boldsymbol{\sigma}}$ projects each speech string in the precover to its shortest prefix that also lies in the precover (see §A for the formal derivation). The conditional probability is then estimated as the ratio

$$\widehat{\overrightarrow{p_{\Sigma}}}_{,N}(\sigma_t \mid \boldsymbol{\sigma}_{<t}) \stackrel{\text{def}}{=} \frac{\widehat{\overrightarrow{p_{\Sigma}}}_{,N}(\boldsymbol{\sigma}_{<t}\sigma_t)}{\widehat{\overrightarrow{p_{\Sigma}}}_{,N}(\boldsymbol{\sigma}_{<t})}. \quad (8)$$

The conditional estimator in Eq. (8) is consistent, meaning it will converge to the true conditional probability as the number of samples approaches infinity (see Proposition 2). We analyze the estimator's variance in §B and the bias of the ratio in Proposition 3.

In our implementation of the estimator, our target distribution is an autoregressive speech LM such as the one described in Eq. (4). We test four speech LMs: a generative speech language model (GSLM) released by Lakhotia et al. (2021), SpiRit-LM (Nguyen et al., 2025), and the small and large variants of Llama-Mimi (Sugiura et al., 2025). See Appendix §C for model details.

As our proposal distribution $q_{\triangle}$, we take a text-to-speech (TTS) model that generates speech for a given input text. Given that TTS models are trained to generate speech corresponding to the input text, we expect that sampling from $q_{\triangle}$ is more likely to yield valid realizations of our target texts than if we were to sample directly from a speech LM. We specifically use Sesame's Conversational Speech Model (CSM),[3] which is trained on multi-speaker conversational speech. We choose a multi-speaker model so that the proposal covers a diverse range of voices. §5.1 analyzes speaker identity in the resulting samples. CSM trains Llama-3.2-1B (Grattafiori et al., 2024) with an alphabet of text and audio tokens, where audio tokens in the alphabet are derived from the first codebook of Mimi speech tokens (Défossez et al., 2024). Mimi is a neural codec trained with both distillation from a semantic audio feature extractor and an autoencoding objective with a discretization bottleneck. The discretized representations at this bottleneck are the speech tokens used in CSM.

For a given string $\boldsymbol{\sigma} \in \Sigma^*$ with $|\boldsymbol{\sigma}| = T$, we sample $N = 500$ spoken realizations from the CSM TTS model that we take as our proposal distribution $q_{\triangle}$. Conditioning on the full text of the stimuli $\boldsymbol{\sigma}$, we generate audio from CSM using a temperature of 1.0. We then use the Montreal Forced Aligner (MFA) (McAuliffe et al., 2017) to segment the sampled audio into realizations of individual orthographic words, giving us N spoken realizations of each of the T units in our string. Using realizations of individual units σ from the same sample $\boldsymbol{\delta}^{(n)}$, we can then calculate $\widehat{\overrightarrow{p_{\Sigma}}}_{,N}(\sigma_t \mid \boldsymbol{\sigma}_{<t})$ using Eq. (8).

The estimator in Eq. (7) requires a prefix monotone mapping function Ω to check for precover membership (see §A). ASR systems that condi-

[3] https://huggingface.co/sesame/csm-1b

tion on the full input sequence do not satisfy this property, as additional audio can revise previously generated text. We therefore require a causal ASR system whose outputs for earlier segments remain fixed as more speech is observed. We approximate this prefix monotone mapping function using the SimulStreaming implementation of Whisper for speech transcription (Radford et al., 2023; Macháček and Polák, 2025). Because our estimator requires transcriptions of all samples and we repeat the estimation for many stimuli, we use Whisper Tiny to balance speed and performance.

3.2 Assessing Surprisal Theory

Using pupillometry data as the psychometric signal of processing effort, we assess the predictive power of surprisal values obtained from our estimates of the marginalized surprisal values for orthographic words from a speech LM. We also compare these against surprisal values derived from unmarginalized orthographic word, phone, and frame probabilities, i.e., the probability the speech LM assigns to units in the specific utterances in the stimuli.

Stimuli In Winn (2016), proportional change in participants' pupil diameters is measured while listening to stimuli and then repeating the stimuli aloud. As in Winn (2016), we use recordings and transcriptions from 162 items in the Revised Speech in Noise corpus (Bilger et al., 1984), each of which corresponds to a single sentence. The stimuli vary in how much the context constrains the final word, for instance ranging from the very constrained phrase "The doctor prescribes the drug" to the more open-ended "Bill might discuss the foam." We only include data from the 21 participants with normal hearing and trials with non-vocoded stimuli in our analysis. We describe our approaches to obtaining surprisal values for the stimuli below.

Estimated Marginalized Surprisal Using the approach to importance sampling discussed in §3.1, we estimate the probability our speech LM assigns to abstract orthographic words in some context, $\widehat{\overrightarrow{p}}_{\Sigma,N}(\sigma_t \mid \boldsymbol{\sigma}_{<t})$. We then obtain surprisal values using the following:

$$\iota_{\text{est}}\left(\sigma_t, \boldsymbol{\sigma}_{<t}\right) = -\log \widehat{\overrightarrow{p}}_{\Sigma,N}\left(\sigma_t \mid \boldsymbol{\sigma}_{<t}\right). \quad (9)$$

Unmarginalized Speech LM Surprisal We compare the estimated marginal surprisals to unmarginalized speech LM surprisal values. For each of the stimuli from Winn (2016), we run the MFA on the original audio to obtain word and phone boundaries. Given these unit boundaries, we obtain conditional probabilities of a specific realization of a unit. We then calculate the surprisal for each unit in the stimuli as follows:

$$\iota_{\text{slm}}\left(\delta_t, \boldsymbol{\delta}_{<t}\right) = -\log \overrightarrow{p}_{\Delta}\left(\delta_t \mid \boldsymbol{\delta}_{<t}\right). \quad (10)$$

We calculate the unmarginalized probabilities of the stimuli at the level of the orthographic word, the phone, and the frame. By frame, we refer to the surprisal for each speech token δ in a string $\boldsymbol{\delta}$, which corresponds to 20 ms of audio for GSLM and 40 ms for SpiRit-LM.

4 Analyses

In the empirical portion of the article, we carry out two analyses. The first serves to confirm that we are sampling utterances from a diverse range of speakers to justify our choice of proposal distribution. Our second analysis assesses surprisal theory using surprisal from speech LMs to model pupil data.

4.1 CSM Sample Diversity

Though the authors anecdotally note variation between the voices and prosodic features of sampled utterances from CSM, we seek to quantitatively assess whether the N realizations of our stimuli indeed capture a wide variety of realizations for the same phrase. To do so, we analyze speaker embeddings of our sampled utterances from a ResNet model pretrained for speaker verification using the SpeechBrain toolkit (Villalba et al., 2020; Ravanelli et al., 2024). By assessing whether we have sampled from a diverse space of speakers, we aim to justify our choice of proposal distribution.

Specifically, we randomly select one of our stimuli and obtain speaker embeddings from each of the N samples of that stimulus obtained from CSM. We then randomly select a speaker from the TIMIT corpus (Garofolo et al., 1993) and obtain speaker embeddings from each of that selected speaker's utterances in the corpus ("within-speaker" embeddings). Finally, we randomly select one of the sentences spoken by all 462 speakers in the TIMIT corpus train split and obtain the speaker embedding for each speaker using that utterance ("between-speaker" embeddings). To assess speaker diversity, we measure the cosine distance between all speaker embeddings in each of the three sets of embeddings.

4.2 Modeling Pupil Size

Generalized additive mixed models (GAMMs) are often used to model psycholinguistic responses given their ability to capture nonlinear relationships between predictors and response variables (van Rij et al., 2019; Berzak and Levy, 2023, *inter alia*). We therefore apply them to compare the predictive power of our surprisal measures for pupillometry data from Winn (2016).

Let y_i denote pupil diameter at time t_i for listener $\ell(i)$ and stimulus $j(i)$. Let $\omega \geq 0$ denote an assumed pupillary lag: a pupil sample recorded at t_i is taken to reflect the linguistic unit active ω ms earlier. Let $m = m(i)$ index the unit active at $t_i - \omega$, and write s_m for its surprisal. We define τ_i as time since onset of that unit, $\tau_i = (t_i - \omega) - \text{onset}(m(i))$.

We run a series of models with two basic structures: a baseline model and a model with current and spillover main effects. Our baseline model consists of random effect smooths for $\ell(i)$ and $j(i)$ to account for listener and stimulus variability. To this baseline, our current and spillover main effects models add main effect smooths for surprisal of the current and K prior units. We fit several main effects models to compare the various marginalized and unmarginalized speech LM surprisals defined in §3.2.

The following is a formal definition of the main effects model described above:

$$\begin{aligned} y_i = &\beta_0 + f_1(s_m) + \sum_{k=1}^{K} f_{1+k}(s_{m-k}) \\ &+ b_{\ell(i)} + b_{j(i)} + \varepsilon_i. \end{aligned} \tag{11}$$

Here, s_{m-k} denotes k-lag spillover over units (e.g., previous words). The functions $f_r(\cdot)$ are univariate smooth effects represented using low-rank penalized regression splines. Random effects b_ℓ and b_j account for listener- and item-level variability, and ε_i represents a Gaussian noise term, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. The baseline includes only random effects and ε_i.

Prior to model fitting, surprisal values are standardized to place predictors on a common scale. In our experiments, we set $K = 1$, corresponding to single-unit spillover terms. Following Winn (2016), we first set $\omega = 1000$ ms. However, as pupillary response is known to vary with factors like luminosity and task difficulty (van Rij et al., 2019), we also test two other physiologically plausible offset times: $\omega \in \{900, 1300\}$. We choose 900 and 1300 ms based on discussion by van Rij et al. (2019).

We fit models with the `bam()` function in the `mgcv` package (Wood, 2006). All smooth terms are estimated using restricted maximum likelihood. Models are compared using ten-fold cross-validation by item (see §D for descriptions of fold size), reporting differences in held-out log-likelihood relative to the baseline. Uncertainty is quantified via bootstrap resampling. To study whether our main effects models result in significant improvements over the baseline model in terms of predicting pupil sizes at a given time, we carry out one-sided paired permutation testing with $B = 5000$ sign-flips and apply the Holm–Bonferroni procedure to correct for conducting multiple significance tests. We also use a one-sided, paired bootstrap test on the difference in held-out delta log-likelihood values to test whether marginalized surprisal values yield significantly higher delta log-likelihoods than unmarginalized surprisal values from the same speech LM.

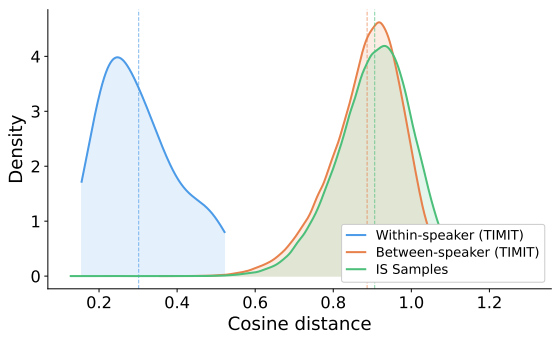

Figure 2: Distributions of cosine distances between speaker embeddings for the within-speaker TIMIT samples, between-speaker TIMIT samples, and the samples generated by CSM for the present study. Dotted vertical lines represent mean cosine distances.

5 Results

5.1 CSM Sample Diversity

When we plot the distributions of cosine distances in our sets of speaker embeddings (Fig. 2), we observe that cosine distances are lower on average in our within-speaker TIMIT sample than in our between-speaker TIMIT sample. The plot shows that the distribution of cosine distances between speaker embeddings from our CSM samples is clearly closer to the between-speaker TIMIT distribution than the within-speaker distribution.

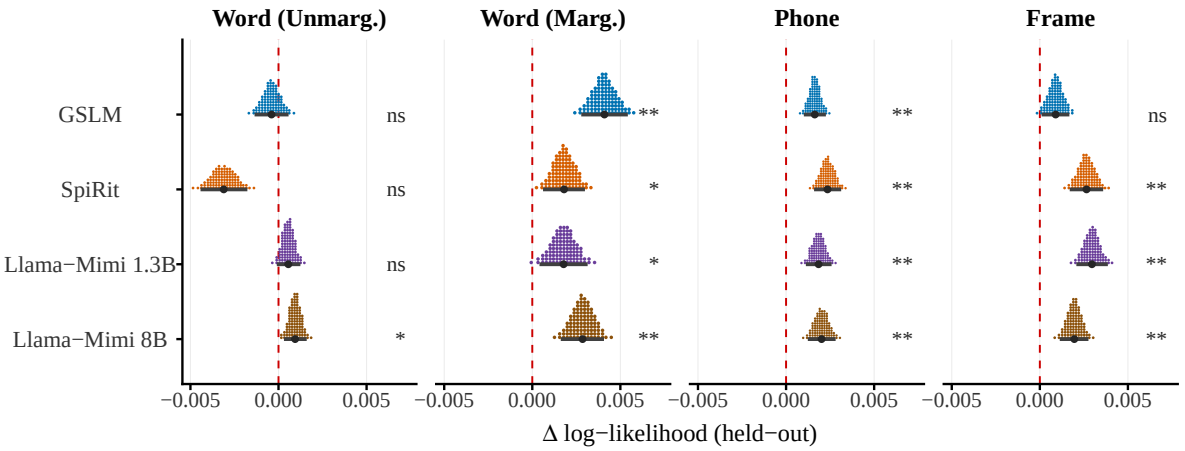

Figure 3: Held-out delta log-likelihood relative to baseline model of pupil size (offset: $\omega = 1000$ ms). Each dot represents 1% of the bootstrap distribution of the mean delta log-likelihood; the point and bars below give the bootstrap mean and 95% CI. Stars correspond to significant corrected p-values of a model compared to the no-surprisal baseline at α levels of 0.05 (*), 0.01 (**), and 0.001 (***); ns = not significant.

5.2 Modeling Pupil Size

Delta log-likelihoods for models of pupil size fit with various surprisal values at offset of $\omega = 1000$ ms are shown in Fig. 3 (see §E for tables of values). Note that delta log-likelihood values can only be compared within each granularity, as each granularity represents a distinct regression problem defined over a different number of pupil observations: orthographic-word windows, phone windows, and frame windows (Kiegeland et al., 2026). At the word level, marginalized and unmarginalized GAMMs within a speech LM are fit on the same folds (and are therefore comparable), but delta log-likelihood is not comparable between speech LMs. Because of estimator error, the ratio in Eq. (8) can yield undefined surprisal values, meaning estimators for different speech LMs successfully yield estimates for different proportions of the tokens in the corpus (coverage for each speech LM is found in §D).

At the level of the orthographic word, we find that for the four speech LMs tested, estimates of the marginalized surprisal values result in larger improvements over the baseline than the unmarginalized values ($p < 0.05$ for all within-model marginalized versus unmarginalized differences; see Tab. 5). Unmarginalized lexical surprisal from speech LMs rarely results in significant improvements over the baseline, with the exception of values obtained from Llama-Mimi 8B.

At the phone granularity, we find that adding unmarginalized phone surprisal consistently yields significant improvements over the baseline. At the level of the frame, unmarginalized surprisal values almost always result in significant improvements over the baseline model, except for GSLM frame surprisals.

We present the mean delta log-likelihoods for analogous models at offsets of $\omega \in \{900, 1000, 1300\}$ ms in Fig. 4 (see §E for tables of values). At the word level, we observe that fitting GAMMs with marginalized surprisals consistently results in higher delta log-likelihoods than fitting GAMMs with unmarginalized lexical surprisals from the same speech LM, regardless of the offset. Improvements in delta log-likelihoods from phone and frame surprisals are generally highest at 900 and 1000 ms offsets.

For reference, we also fit models of pupil size using word-level surprisals obtained from the text LM GPT-2 (Radford et al., 2019) over the stimuli transcriptions (see §E.1 for details and results). These also result in significant improvements over the baseline with no surprisal. The larger the offset used, the less predictive GPT-2-derived surprisals are of pupil size. We note that comparison with word-level delta log-likelihoods from speech LMs is not warranted. Regardless of the specific delta log-likelihoods obtained, our approach derives the target predictor (surprisal) from the same input signal as participants in the pupillometry study, making it a more valid approach to obtaining surprisal in a listening comprehension study than using the transcribed speech as the input signal.

Finally, we refit our models, now adding a smooth term for the item-normalized time at which pupil dilation is recorded. Adding this term to both the main effects models and the baseline allows us to consider the extent to which surprisal adds predictive power to pupil dilation beyond the time in the sequence. The results indicate that adding surprisal to the baseline of time and random effects for participants and items does not result in significant improvements, regardless of the modality or granularity of the unit.

6 Discussion

In the present work, we asked how surprisal should be defined when it is estimated from the acoustic signal using a speech LM rather than from text. Because speech LMs assign probabilities to speech-code sequences, their probabilities can be interpreted either as probabilities of specific acoustic realizations or, after marginalizing over possible realizations, as probabilities of abstract linguistic units. Our results favor the latter interpretation at the word level: marginalized surprisal estimates for abstract lexical units were more predictive of pupil dilation than realization-specific word surprisal. At the same time, surprisal computed over finer-grained units (phones and frames) also improved models of pupil size, suggesting that speech-LM probabilities can model information relevant to listening effort at multiple granularities.

Taken together, our results suggest that in speech processing, humans are able to abstract over varying acoustic characteristics of input and use abstract lexical units to make inferences about upcoming words. This being the case, models of speech comprehension should distinguish between the probability of a particular acoustic realization and the probability of the abstract lexical unit it realizes. This is consistent with the speech perception literature: ERP data collected while participants listened to stories indicate that listeners anticipate next words given a context (Van Berkum et al., 2005), and word recognition in noisy conditions is easier when the word is semantically or grammatically predictable in context (Miller and Isard, 1963). Moreover, there is evidence from pupillometry studies that with more context, repairing misheard words requires less cognitive effort (Smith and Winn, 2025). Our findings also parallel work on reading comprehension, for example the finding by Vigly et al. (2025) that surprisal of surface forms is no longer predictive of reading times in texts with typographical errors.

Our findings for phones and frames are consistent with surprisal theory research that has previously postulated that sub-lexical units are also processed in a predictive manner (Smith and Levy, 2013). In certain contexts, it has been shown that finer-grained units may play a role in probabilistic anticipation during speech processing. For example, eye-tracking data indicates that variation in voice onset time, a primary correlate to the stop voicing contrast (e.g., /p/ vs. /b/) may be used to predict an upcoming word in ambiguous contexts (McMurray et al., 2009). Given that predictions from cues like voice onset time are based on even finer-grained input than the phone level, such anticipatory processing may explain the significance of our frame-level surprisal models, despite these units not being linguistically motivated.

Our findings show that surprisal from speech LMs varies in how well it predicts pupil dilation depending on granularity: for example, whereas unmarginalized phone surprisals from GSLM result in significant improvements over the baseline model, GSLM's unmarginalized frame and word surprisals do not. While there is a large body of work studying how characteristics such as size and training data impact the so-called psychometric predictive power (PPP) of text LMs (Kuribayashi et al., 2024; Oh et al., 2025, *inter alia*), this line of work has yet to be explored for speech LMs. Modeling decisions that yield high PPP for some units may not necessarily do the same for other units, so future work should study how different speech LM characteristics impact their cognitive plausibility across unit inventories.

Our finding that surprisal—regardless of modality, abstraction, or unit over which it is obtained—does not add predictive value over a baseline that includes information about when a pupil measurement was taken is consistent with a growing body of literature about the relative importance of surprisal as a predictor in sentence processing. Indeed, it has been shown that factors such as unigram frequency, word length, oculomotor noise, and parafoveal preview can all explain more variance in reading times than surprisal alone (Heilbron et al., 2023; Opedal et al., 2024). Thus, while we provide an approach to estimating word probabilities from speech LMs, our empirical results add to the body of work suggesting a relatively less important role for surprisal in sentence processing.

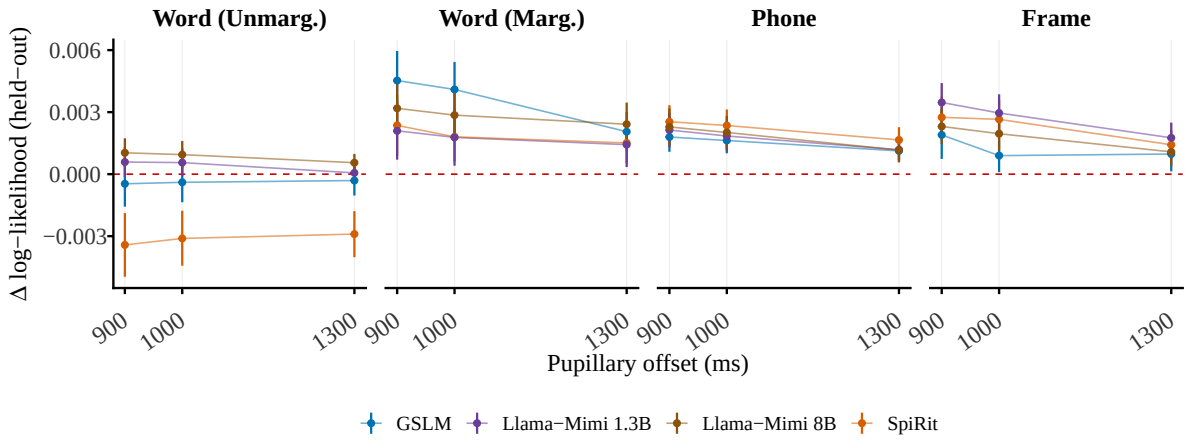

Figure 4: Held-out delta log-likelihood relative to baseline model of pupil size at various offset times. Points give the mean delta log-likelihood and bars give 95% CI. Values at 1000 ms reflect those in Fig. 3. See §E for significance of delta log-likelihood values over the baseline model.

Our results also have implications for natural language processing. Notably, they call into question the construct validity of common speech language modeling benchmarks such as sBLIMP and sWUGGY (Nguyen et al., 2020). In these benchmarks, recordings of words and non-words (sWUGGY) or syntactically correct and incorrect (sBLIMP) sentences are given to speech language models, and higher scores are given when models deem the words and grammatical audio more probable than the non-words and ungrammatical audio. That is, these benchmarks score models based on the probability they assign to individual realizations of words or sentences. As we have argued here, the probabilities assigned by speech LMs to specific utterances are not representative of the probabilities the model assigns to abstract linguistic units, as the probabilities can reflect the specific utterance's prosodic, acoustic, or speaker characteristics. This is particularly concerning for cases where the benchmark has been applied to evaluate speech LMs that presume to model abstract, invariant units (Poli et al., 2025; Chang et al., 2025, *inter alia*). We hope our work encourages further benchmark development for speech language modeling that accounts for the idea that the relative probability assigned to a specific utterance might change if we manipulate the prosody, voice, or other properties of the audio.

7 Conclusion

The current body of literature on surprisal theory applies mainly to questions of sentence processing during reading rather than spoken language comprehension. Though speech language models provide a means of applying a similar research program to spoken language comprehension, obtaining surprisal from the speech signal brings several added challenges: namely, the specific psychometric signal to model, the question of which units to consider given the naturally continuous nature of speech, and the complication of modeling many variations of the same abstract unit. We address the last of these with an importance sampling estimator that approximates the surprisal a speech LM assigns to abstract units. Among the approaches to obtaining surprisal from a speech LM that were tested, we found that our estimations of abstract, lexical surprisals better modeled our signal of processing effort, pupil size, than realization-specific word probabilities. We also found that phone and frame surprisal from speech LMs improved models of pupil size. These findings suggest that the surprisal of abstract lexical units is a key component of processing difficulty in spoken language comprehension. Our approach to estimating surprisal for abstract units opens the door to further study the contexts in which we base our predictions of upcoming input on sub-lexical units and specific realizations.

Limitations

Our experiments focus on a single listening study (Winn, 2016) with 162 items from the Revised Speech in Noise corpus (Bilger et al., 1984). We only analyze data from normal hearing listeners,

but further examination would be needed to draw conclusions about whether the results hold for people with hearing loss, who have been shown to use context differently than normal hearing listeners (Winn, 2016). The nature of the stimuli limits the extent to which we can study sub-lexical phenomena and the impact of prosody and indexical information in the speech signal on processing effort during listening. In stimuli designed to study listening in unfamiliar stress patterns (e.g., accented stress patterns), the relative predictive power of marginalized and unmarginalized surprisals from a speech LM may differ. Likewise, we could imagine the following stimuli in a listening study: "I reached into my toolbox and grabbed a hamburger." The word *hamburger* is surprising in this context, but because the phonetic sequence /hæm-/ is expected (e.g., as the start of the word "hammer"), phone surprisal might also be predictive of pupil size. That being the case, without stimuli designed to capture sub-lexical effects, we likely do not fully characterize the role of finer-grained units' surprisal during speech processing. We are also limited to read (rather than spontaneous) speech in English in this study.

Another limitation of the present study is the ASR sample rejection; though this step is necessary to exclude TTS errors, it is also a point at which errors can be made. Though a fuller error analysis is not possible because we do not have ground truth transcriptions for the sampled audio, we suspect that the most likely errors would be mis-transcriptions of valid realizations of a word. This may result in a bias against audio with especially atypical prosody or speaker features. We expect that exploring larger ASR models would decrease such errors, albeit at higher computational costs. We expect that better ASR models and a higher initial sample size N would also contribute to higher estimator coverage overall.

We also note that the present work only studies abstract units at the level of orthographic words and does not estimate marginalized phone surprisals. A direct application of our importance sampling approach to the phone granularity would yield estimates of the probabilities speech LMs assign to maximally context-sensitive phones in our stimuli across many speakers. However, extending our approach to study less context-sensitive units at the phone level (e.g., allophones or phonemes) would require extending the method to also sample the relevant units in diverse phonological contexts.

Ethical Concerns

We use only previously studied pupillometry and speech data with explicit permission from the author (Winn, 2016). We do not foresee any ethical concerns from the use of this data to study human speech processing.

Acknowledgments

ZH was supported by the Swiss National Science Foundation (Grant No. 10003607).

References

Siddhant Arora, Kai-Wei Chang, Chung-Ming Chien, Yifan Peng, Haibin Wu, Yossi Adi, Emmanuel Dupoux, Hung-Yi Lee, Karen Livescu, and Shinji Watanabe. 2025. On the landscape of spoken language models: A comprehensive survey. *Transactions on Machine Learning Research*.

Yevgeni Berzak and Roger Levy. 2023. Eye movement traces of linguistic knowledge in native and non-native reading. *Open Mind*, 7.

Robert C. Bilger, Jakob M. Nuetzel, William M. Rabinowitz, and C. Rzeczkowski. 1984. Standardization of a test of speech perception in noise. *Journal of Speech, Language, and Hearing Research*, 27(1).

Heng-Jui Chang, Hongyu Gong, Changhan Wang, James Glass, and Yu-An Chung. 2025. DC-Spin: A speaker-invariant speech tokenizer for spoken language models. In *Proceedings of Interspeech 2025*.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: A speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*.

John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, and David S. Pallett. 1993. DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon Technical Report N*, 93.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

H. O. Hartley and A. Ross. 1954. Unbiased ratio estimators. *Nature*, 174(4423).

Micha Heilbron, Jorie van Haren, Peter Hagoort, and Floris P. de Lange. 2023. Lexical processing strongly affects reading times but not skipping during natural reading. *Open Mind*, 7.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE ACM Transactions on Audio, Speech, and Language Processing*, 29.

Samuel Kiegeland, 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)*.

Tatsuki Kuribayashi, Yohei Oseki, and Timothy Baldwin. 2024. Psychometric predictive power of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9.

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

Alvin M. Liberman, Franklin S. Cooper, Donald P. Shankweiler, and Michael Studdert-Kennedy. 1967. Perception of the speech code. *Psychological Review*, 74(6).

Dominik Macháček and Peter Polák. 2025. Simultaneous translation with offline speech and LLM models in CUNI submission to IWSLT 2025. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In *Proceedings of Interspeech 2017*.

Bob McMurray, Michael K. Tanenhaus, and Richard N. Aslin. 2009. Within-category VOT affects recovery from "lexical" garden-paths: Evidence against phoneme-level inhibition. *Journal of Memory and Language*, 60(1).

George A. Miller and Stephen Isard. 1963. Some perceptual consequences of linguistic rules. *Journal of Verbal Learning and Verbal Behavior*, 2(3).

Tu Anh Nguyen, Maureen de Seyssel, Patricia Rozé, Morgane Rivière, Evgeny Kharitonov, Alexei Baevski, Ewan Dunbar, and Emmanuel Dupoux. 2020. The Zero Resource Speech Benchmark 2021: Metrics and baselines for unsupervised spoken language modeling. In *NeuRIPS Workshop on Self-Supervised Learning for Speech and Audio Processing*.

Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R. Costa-jussa, Maha Elbayad, Sravya Popuri, Christophe Ropers, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, Itai Gat, Mary Williamson, Gabriel Synnaeve, Juan Pino, Benoît Sagot, and Emmanuel Dupoux. 2025. SpiRit-LM: Interleaved spoken and written language model. *Transactions of the Association for Computational Linguistics*, 13.

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 Wilcox. 2024. On the role of context in reading time prediction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Gordon E. Peterson and Harold L. Barney. 1952. Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, 24(2).

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

Maxime Poli, Mahi Luthra, Youssef Benchekroun, Yosuke Higuchi, Martin Gleize, Jiayi Shen, Robin Algayres, Yu-An Chung, Mido Assran, Juan Pino, and Emmanuel Dupoux. 2025. SpidR: Learning fast and stable linguistic units for spoken language models without supervision. *Transactions on Machine Learning Research*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the International Conference on Machine Learning*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8).

Mirco Ravanelli, Titouan Parcollet, Adel Moumen, Sylvain de Langen, Cem Subakan, Peter Plantinga, Yingzhi Wang, Pooneh Mousavi, Luca Della Libera, Artem Ploujnikov, Francesco Paissan, Davide Borra, Salah Zaiem, Zeyu Zhao, Shucong Zhang, Georgios Karakasidis, Sung-Lin Yeh, Pierre Champion, Aku Rouhe, and 14 others. 2024. Open-source conversational AI with SpeechBrain 1.0. *Journal of Machine Learning Research*, 25(333).

Christian P. Robert and George Casella. 2004. *Monte Carlo Statistical Methods*, volume 2. Springer.

Wendy Sandler and Diane Lillo-Martin. 2006. *Sign Language and Linguistic Universals*. Cambridge University Press.

Cory Shain, Idan Asher Blank, Marten van Schijndel, William Schuler, and Evelina Fedorenko. 2020. fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, 138.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. 2024. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10).

Michael L. Smith and Matthew B. Winn. 2025. Repairing misperceptions of words early in a sentence is more effortful than repairing later words, especially for listeners with cochlear implants. *Trends in Hearing*, 29.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3).

Vésteinn Snæbjarnarson, Samuel Kiegeland, Tianyu Liu, Reda Boumasmoud, Tim Vieira, and Ryan Cotterell. 2026. Transducing language models. In *The International Conference on Learning Representations*.

Ediz Sohoglu, Loes Beckers, and Matthew H. Davis. 2024. Convergent neural signatures of speech prediction error are a biological marker for spoken word recognition. *Nature Communications*, 15(1).

Issa Sugiura, Shuhei Kurita, Yusuke Oda, and Ryuichiro Higashinaka. 2025. Llama-Mimi: Exploring the limits of flattened speech language modeling. *arXiv preprint arXiv:2509.14882*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jos J. A. Van Berkum, Colin M. Brown, Pienie Zwitserlood, Valesca Kooijman, and Peter Hagoort. 2005. Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(3).

Jacolien van Rij, Petra Hendriks, Hedderik van Rijn, R. Harald Baayen, and Simon N. Wood. 2019. Analyzing the time course of pupillometric data. *Trends in Hearing*, 23.

Tim Vieira, Benjamin LeBrun, Mario Giulianelli, Juan Luis Gastaldi, Brian DuSell, John Terilla, Timothy J. O'Donnell, and Ryan Cotterell. 2025. From language models over tokens to language models over characters. In *Proceedings of the International Conference on Machine Learning*.

Jacob Vigly, Peng Qian, Morgan Sonderegger, and Timothy O'Donnell. 2025. Comprehension effort as the cost of inference. *PsyArXiv*.

Jesús Villalba, Nanxin Chen, David Snyder, Daniel Garcia-Romero, Alan McCree, Gregory Sell, Jonas Borgstrom, Leibny Paola García-Perera, Fred Richardson, Réda Dehak, Pedro A. Torres-Carrasquillo, and Najim Dehak. 2020. State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and speakers in the wild evaluations. *Computer Speech & Language*, 60.

Sarenne Carrol Wallbridge, Catherine Lai, and Peter Bell. 2022. Investigating perception of spoken dialogue acceptability through surprisal. In *Proceedings of Interspeech 2022*.

Matthew B. Winn. 2016. Rapid release from listening effort resulting from semantic context, and effects of spectral degradation and cochlear implants. *Trends in Hearing*, 20.

S.N. Wood. 2006. *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC.

Appendix Contents

A Derivation of the Importance Sampling Estimator **14**

B Estimator Variance **16**

C Speech LM Details **16**

D GAMM Data Summary **17**

E Supplementary Results **17**

E.1 GPT-2 Results 18

E.2 Tables of Values 18

E.3 Marginalized and Unmarginalized Comparisons 18

E.4 Pupil Time Baseline 18

A Derivation of the Importance Sampling Estimator

Proposition 1. *Let $p_\triangle \ll q_\triangle$ (i.e., $p_\triangle$ is absolutely continuous with respect to $q_\triangle$), and let $\Omega\colon \triangle^* \to \Sigma^*$ be a prefix monotone function. Define the precover $\mathcal{P}(\boldsymbol{\sigma}) \stackrel{\text{def}}{=} \{\boldsymbol{\delta} \in \triangle^* \colon \Omega(\boldsymbol{\delta}) \succeq \boldsymbol{\sigma}\}$. Let $\pi_{\boldsymbol{\sigma}}\colon \mathcal{P}(\boldsymbol{\sigma}) \to \mathrm{pf}(\mathcal{P}(\boldsymbol{\sigma}))$ project each $\boldsymbol{\delta}$ to its longest prefix in $\mathrm{pf}(\mathcal{P}(\boldsymbol{\sigma}))$.[4] Then, for any $\boldsymbol{\sigma} \in \Sigma^*$, we have*

$$\overrightarrow{p_\Sigma}(\boldsymbol{\sigma}) = \mathop{\mathbb{E}}_{\boldsymbol{\delta}\sim q_\triangle} \mathbb{1}\{\boldsymbol{\delta} \in \mathcal{P}(\boldsymbol{\sigma})\} \frac{\overrightarrow{p_\triangle}(\pi_{\boldsymbol{\sigma}}(\boldsymbol{\delta}))}{\overrightarrow{q_\triangle}(\pi_{\boldsymbol{\sigma}}(\boldsymbol{\delta}))} \tag{12}$$

Proof.

$$\begin{aligned}
\overrightarrow{p_\Sigma}(\boldsymbol{\sigma}) &= \sum_{\boldsymbol{\delta}\in\triangle^*} \mathbb{1}\{\Omega(\boldsymbol{\delta}) \succeq \boldsymbol{\sigma}\} p_\triangle(\boldsymbol{\delta}) && \text{(13a, def. of } \overrightarrow{p_\Sigma}\text{)}\\
&= \sum_{\boldsymbol{\delta}\in\mathcal{P}(\boldsymbol{\sigma})} p_\triangle(\boldsymbol{\delta}) && \text{(13b, def. of } \mathcal{P}(\boldsymbol{\sigma})\text{)}\\
&= \sum_{\boldsymbol{\delta}\in\mathrm{pf}(\mathcal{P}(\boldsymbol{\sigma}))} \overrightarrow{p_\triangle}(\boldsymbol{\delta}) && \text{(13c, } \mathcal{P}(\boldsymbol{\sigma}) \text{ decomposes into disjoint cylinders)}\\
&= \sum_{\boldsymbol{\delta}\in\mathrm{pf}(\mathcal{P}(\boldsymbol{\sigma}))} \overrightarrow{q_\triangle}(\boldsymbol{\delta}) \frac{\overrightarrow{p_\triangle}(\boldsymbol{\delta})}{\overrightarrow{q_\triangle}(\boldsymbol{\delta})} && \text{(13d, multiply and divide by } \overrightarrow{q_\triangle}\text{; valid since } p_\triangle \ll q_\triangle\text{)}\\
&= \sum_{\boldsymbol{\delta}\in\triangle^*} q_\triangle(\boldsymbol{\delta}) \mathbb{1}\{\boldsymbol{\delta} \in \mathcal{P}(\boldsymbol{\sigma})\} \frac{\overrightarrow{p_\triangle}(\pi_{\boldsymbol{\sigma}}(\boldsymbol{\delta}))}{\overrightarrow{q_\triangle}(\pi_{\boldsymbol{\sigma}}(\boldsymbol{\delta}))} && \text{(13e, } \textstyle\sum_{\boldsymbol{\delta}'\succeq\boldsymbol{\delta}} q_\triangle(\boldsymbol{\delta}') = \overrightarrow{q_\triangle}(\boldsymbol{\delta})\text{; reindex via } \pi_{\boldsymbol{\sigma}}\text{)}\\
&= \mathop{\mathbb{E}}_{\boldsymbol{\delta}\sim q_\triangle} \mathbb{1}\{\boldsymbol{\delta} \in \mathcal{P}(\boldsymbol{\sigma})\} \frac{\overrightarrow{p_\triangle}(\pi_{\boldsymbol{\sigma}}(\boldsymbol{\delta}))}{\overrightarrow{q_\triangle}(\pi_{\boldsymbol{\sigma}}(\boldsymbol{\delta}))}. && \text{(13f, def. of expectation under } q_\triangle\text{)}
\end{aligned}$$

■

Given N i.i.d. samples $\boldsymbol{\delta}^{(1)}, \ldots, \boldsymbol{\delta}^{(N)} \sim q_\triangle$, the Monte Carlo estimator corresponding to Eq. (12) is

$$\widehat{\overrightarrow{p_\Sigma}}_N(\boldsymbol{\sigma}) \stackrel{\text{def}}{=} \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\left\{\boldsymbol{\delta}^{(n)} \in \mathcal{P}(\boldsymbol{\sigma})\right\} \frac{\overrightarrow{p_\triangle}(\pi_{\boldsymbol{\sigma}}(\boldsymbol{\delta}^{(n)}))}{\overrightarrow{q_\triangle}(\pi_{\boldsymbol{\sigma}}(\boldsymbol{\delta}^{(n)}))}. \tag{14}$$

The conditional prefix probability is then estimated as the ratio

$$\widehat{\overrightarrow{p_\Sigma}}_N(\sigma_t \mid \boldsymbol{\sigma}_{<t}) \stackrel{\text{def}}{=} \frac{\widehat{\overrightarrow{p_\Sigma}}_N(\boldsymbol{\sigma}_{<t}\sigma_t)}{\widehat{\overrightarrow{p_\Sigma}}_N(\boldsymbol{\sigma}_{<t})}. \tag{15}$$

Proposition 2. *Let $\boldsymbol{\delta}^{(1)}, \ldots \boldsymbol{\delta}^{(N)} \overset{\text{i.i.d.}}{\sim} q_\triangle$ and suppose that*

$$\mathop{\mathbb{E}}_{\boldsymbol{\delta}\sim q_\triangle} \left[\left(\mathbb{1}\{\boldsymbol{\delta} \in \mathcal{P}(\boldsymbol{\sigma})\} \frac{\overrightarrow{p_\triangle}(\pi_{\boldsymbol{\sigma}}(\boldsymbol{\delta}))}{\overrightarrow{q_\triangle}(\pi_{\boldsymbol{\sigma}}(\boldsymbol{\delta}))} \right)^2 \right] < \infty \tag{16}$$

for all $\boldsymbol{\sigma} \in \Sigma^*$ and that $\overrightarrow{p_\Sigma}(\boldsymbol{\sigma}) > 0$ for all $\boldsymbol{\sigma} \in \Sigma^*$. Then, we have that

$$\widehat{\overrightarrow{p_\Sigma}}_N(\sigma_t \mid \boldsymbol{\sigma}_{<t}) \xrightarrow[N\to\infty]{\text{a.s.}} \overrightarrow{p_\Sigma}(\sigma_t \mid \boldsymbol{\sigma}_{<t}). \tag{17}$$

Proof. The integrand in Eq. (14) has finite second moment by assumption. By the strong law of large numbers, $\widehat{\overrightarrow{p_\Sigma}}_N(\boldsymbol{\sigma}) \xrightarrow[N\to\infty]{\text{a.s.}} \overrightarrow{p_\Sigma}(\boldsymbol{\sigma})$ for all $\boldsymbol{\sigma} \in \Sigma^*$. Because $\overrightarrow{p_\Sigma}(\boldsymbol{\sigma}_{<t}) > 0$, the function $f(x, y) = x/y$ is continuous at $(\overrightarrow{p_\Sigma}(\boldsymbol{\sigma}_{<t}\sigma_t), \overrightarrow{p_\Sigma}(\boldsymbol{\sigma}_{<t}))$. By the continuous mapping theorem, we arrive at

$$\widehat{\overrightarrow{p_\Sigma}}_N(\sigma_t \mid \boldsymbol{\sigma}_{<t}) = f(\widehat{\overrightarrow{p_\Sigma}}_N(\boldsymbol{\sigma}_{<t}\sigma_t), \widehat{\overrightarrow{p_\Sigma}}_N(\boldsymbol{\sigma}_{<t})) \xrightarrow[N\to\infty]{\text{a.s.}} f(\overrightarrow{p_\Sigma}(\boldsymbol{\sigma}_{<t}\sigma_t), \overrightarrow{p_\Sigma}(\boldsymbol{\sigma}_{<t})) = \overrightarrow{p_\Sigma}(\sigma_t \mid \boldsymbol{\sigma}_{<t}). \tag{18}$$

■

[4] The prefix base of the precover is defined as $\mathrm{pf}(\mathcal{P}(\boldsymbol{\sigma})) \stackrel{\text{def}}{=} \{\boldsymbol{\delta} \in \mathcal{P}(\boldsymbol{\sigma}) : \nexists\, \boldsymbol{\delta}' \in \mathcal{P}(\boldsymbol{\sigma}) \text{ s.t. } \boldsymbol{\delta}' \prec \boldsymbol{\delta}\}$.

Proposition 3. *Adopt the assumptions of Proposition 1, and suppose $\overrightarrow{p_\Sigma}(\boldsymbol{\sigma}_{<t}) > 0$. For $\delta \sim q_\triangle$ define the single-sample importance weights*

$$A \stackrel{\text{def}}{=} \mathbb{1}\{\delta \in \mathcal{P}(\boldsymbol{\sigma}_{<t}\sigma_t)\}\frac{\overrightarrow{p_\triangle}(\pi_{\boldsymbol{\sigma}_{<t}\sigma_t}(\delta))}{\overrightarrow{q_\triangle}(\pi_{\boldsymbol{\sigma}_{<t}\sigma_t}(\delta))},$$
$$B \stackrel{\text{def}}{=} \mathbb{1}\{\delta \in \mathcal{P}(\boldsymbol{\sigma}_{<t})\}\frac{\overrightarrow{p_\triangle}(\pi_{\boldsymbol{\sigma}_{<t}}(\delta))}{\overrightarrow{q_\triangle}(\pi_{\boldsymbol{\sigma}_{<t}}(\delta))},$$

and suppose $\mathbb{E}[A^2] < \infty$ and $\mathbb{E}[B^2] < \infty$. Assume further that the denominator estimator $\widehat{\overrightarrow{p_\Sigma}}_{,N}(\boldsymbol{\sigma}_{<t})$ is almost surely positive and that the conditional estimator from Eq. (15) has finite variance. Writing its bias as

$$\mathrm{Bias}_N \stackrel{\text{def}}{=} \mathbb{E}\left[\widehat{\overrightarrow{p_\Sigma}}_{,N}(\sigma_t \mid \boldsymbol{\sigma}_{<t})\right] - \overrightarrow{p_\Sigma}(\sigma_t \mid \boldsymbol{\sigma}_{<t}),$$

the following hold.

(i) (Exact bias identity.)

$$\mathrm{Bias}_N = -\frac{\mathrm{Cov}\left(\widehat{\overrightarrow{p_\Sigma}}_{,N}(\sigma_t \mid \boldsymbol{\sigma}_{<t}),\ \widehat{\overrightarrow{p_\Sigma}}_{,N}(\boldsymbol{\sigma}_{<t})\right)}{\overrightarrow{p_\Sigma}(\boldsymbol{\sigma}_{<t})}. \tag{19}$$

(ii) (Relative-bias bound.)

$$\frac{|\mathrm{Bias}_N|}{\sqrt{\mathbb{V}\left(\widehat{\overrightarrow{p_\Sigma}}_{,N}(\sigma_t \mid \boldsymbol{\sigma}_{<t})\right)}} \le \frac{\sqrt{\mathbb{V}\left(\widehat{\overrightarrow{p_\Sigma}}_{,N}(\boldsymbol{\sigma}_{<t})\right)}}{\overrightarrow{p_\Sigma}(\boldsymbol{\sigma}_{<t})} = \frac{1}{\sqrt{N}}\frac{\sqrt{\mathbb{V}(B)}}{\overrightarrow{p_\Sigma}(\boldsymbol{\sigma}_{<t})}. \tag{20}$$

In particular, the bias measured relative to the estimator's own standard error is $O(N^{-1/2})$: it shrinks strictly faster than the sampling noise and is thus asymptotically negligible relative to it.

Proof. Write $R \stackrel{\text{def}}{=} \widehat{\overrightarrow{p_\Sigma}}_{,N}(\sigma_t \mid \boldsymbol{\sigma}_{<t}) = \widehat{\overrightarrow{p_\Sigma}}_{,N}(\boldsymbol{\sigma}_{<t}\sigma_t)\,/\,\widehat{\overrightarrow{p_\Sigma}}_{,N}(\boldsymbol{\sigma}_{<t})$ and $r \stackrel{\text{def}}{=} \overrightarrow{p_\Sigma}(\sigma_t \mid \boldsymbol{\sigma}_{<t}) = \overrightarrow{p_\Sigma}(\boldsymbol{\sigma}_{<t}\sigma_t)/\overrightarrow{p_\Sigma}(\boldsymbol{\sigma}_{<t})$. Each of $\widehat{\overrightarrow{p_\Sigma}}_{,N}(\boldsymbol{\sigma}_{<t}\sigma_t)$ and $\widehat{\overrightarrow{p_\Sigma}}_{,N}(\boldsymbol{\sigma}_{<t})$ is a sample mean of N i.i.d. copies of A and B respectively (cf. Eq. (14)), so by Proposition 1, $\mathbb{E}[\widehat{\overrightarrow{p_\Sigma}}_{,N}(\boldsymbol{\sigma}_{<t}\sigma_t)] = \overrightarrow{p_\Sigma}(\boldsymbol{\sigma}_{<t}\sigma_t)$ and $\mathbb{E}[\widehat{\overrightarrow{p_\Sigma}}_{,N}(\boldsymbol{\sigma}_{<t})] = \overrightarrow{p_\Sigma}(\boldsymbol{\sigma}_{<t})$.

Part (i). Because $\widehat{\overrightarrow{p_\Sigma}}_{,N}(\boldsymbol{\sigma}_{<t}) > 0$ almost surely, $R \cdot \widehat{\overrightarrow{p_\Sigma}}_{,N}(\boldsymbol{\sigma}_{<t}) = \widehat{\overrightarrow{p_\Sigma}}_{,N}(\boldsymbol{\sigma}_{<t}\sigma_t)$ identically, and therefore

$$\begin{aligned}\mathrm{Cov}\left(R,\ \widehat{\overrightarrow{p_\Sigma}}_{,N}(\boldsymbol{\sigma}_{<t})\right) &= \mathbb{E}\left[R\,\widehat{\overrightarrow{p_\Sigma}}_{,N}(\boldsymbol{\sigma}_{<t})\right] - \mathbb{E}[R]\ \mathbb{E}\left[\widehat{\overrightarrow{p_\Sigma}}_{,N}(\boldsymbol{\sigma}_{<t})\right] \\ &= \mathbb{E}\left[\widehat{\overrightarrow{p_\Sigma}}_{,N}(\boldsymbol{\sigma}_{<t}\sigma_t)\right] - \mathbb{E}[R]\,\overrightarrow{p_\Sigma}(\boldsymbol{\sigma}_{<t}) \\ &= \overrightarrow{p_\Sigma}(\boldsymbol{\sigma}_{<t}\sigma_t) - \mathbb{E}[R]\,\overrightarrow{p_\Sigma}(\boldsymbol{\sigma}_{<t}).\end{aligned} \tag{21}$$

Dividing Eq. (21) by $\overrightarrow{p_\Sigma}(\boldsymbol{\sigma}_{<t}) > 0$ and rearranging,

$$\mathbb{E}[R] - \underbrace{\frac{\overrightarrow{p_\Sigma}(\boldsymbol{\sigma}_{<t}\sigma_t)}{\overrightarrow{p_\Sigma}(\boldsymbol{\sigma}_{<t})}}_{=r} = -\frac{\mathrm{Cov}\left(R,\ \widehat{\overrightarrow{p_\Sigma}}_{,N}(\boldsymbol{\sigma}_{<t})\right)}{\overrightarrow{p_\Sigma}(\boldsymbol{\sigma}_{<t})},$$

which is Eq. (19).

Part (ii). By the Cauchy–Schwarz inequality, $\left|\mathrm{Cov}\left(R,\ \widehat{\overrightarrow{p_\Sigma}}_{,N}(\boldsymbol{\sigma}_{<t})\right)\right| \le \sqrt{\mathbb{V}(R)\,\mathbb{V}\left(\widehat{\overrightarrow{p_\Sigma}}_{,N}(\boldsymbol{\sigma}_{<t})\right)}$. Substituting into Eq. (19) and dividing by $\sqrt{\mathbb{V}(R)}$ gives the bound in Eq. (20); the final equality uses $\mathbb{V}\left(\widehat{\overrightarrow{p_\Sigma}}_{,N}(\boldsymbol{\sigma}_{<t})\right) = \mathbb{V}(B)/N$, as $\widehat{\overrightarrow{p_\Sigma}}_{,N}(\boldsymbol{\sigma}_{<t})$ averages N i.i.d. copies of B. Since the right-hand side of Eq. (20) is $O(N^{-1/2})$, the relative bias vanishes at rate $N^{-1/2}$. The inequality in Eq. (20) is the classical relative-bias bound for a ratio estimator (Hartley and Ross, 1954). ■

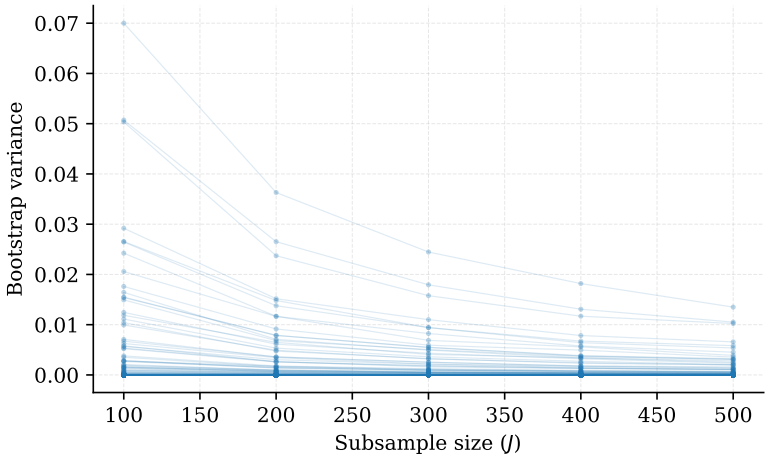

Figure 5: Variance of the IS estimator estimated via nonparametric bootstrap (1000 resamples per subsample size). Target speech LM is Llama-Mimi 1.3B. Each point shows the bootstrap variance for a given word at a given subsample size, and lines connect the same word across sample sizes. The top 2% of variances are excluded for visualization.

B Estimator Variance

As a diagnostic of the stability of our implementation of the estimator in Eq. (7), we analyze its sampling variance. This is akin to asking how much our estimate would change if we drew a different set of samples from CSM. If we are sampling from the full possible acoustic space of an abstracted unit, the specific set of samples should not impact the final estimate, meaning the variance should be low when we resample. We assess this by repeatedly resampling from our data and measuring how much the estimates change. For our estimator $\widehat{\overrightarrow{p_{\Sigma}}}_{,N}$ computed from N samples, the sampling variance is defined as

$$\mathbb{V}\left(\widehat{\overrightarrow{p_{\Sigma}}}_{,N}\right) = \mathbb{E}\left[(\widehat{\overrightarrow{p_{\Sigma}}}_{,N} - \mathbb{E}\left[\widehat{\overrightarrow{p_{\Sigma}}}_{,N}\right])^2\right]. \tag{22}$$

Because generating independent Monte Carlo runs is computationally expensive, we obtain a single set of 500 importance samples per orthographic word in the Winn (2016) stimuli. To approximate the sampling variance at different effective sample sizes, we employ nonparametric bootstrap resampling.

For each word and for each subsample $J \in \{100, 200, 300, 400, 500\}$, we draw 1,000 bootstrap samples (with replacement) of size J from the original 500 importance weights and recompute the estimator. The variance across these bootstrap replicates provides an estimate of $\mathbb{V}\left(\widehat{\overrightarrow{p_{\Sigma}}}_{,J}\right)$. We find that across orthographic words in our stimuli, our estimator yields lower variance estimates of the word's probability with larger sample sizes (see Fig. 5). That is, as the sample size used to estimate a word's probability increases, the estimation becomes more robust to the specific sample used.

C Speech LM Details

Here we give further details on the speech LMs we use as target distributions $p_{\triangle}$ to obtain surprisal values from the speech signal. The generative speech language model (GSLM) released by Lakhotia et al. (2021) is a transformer decoder trained with a context length of 3,072 speech tokens. The speech tokens are drawn from a 200-unit codebook obtained by applying k-means clustering to features from layer 6 of the HuBERT feature extractor (Hsu et al., 2021). Due to the temporal resolution of HuBERT, each speech token corresponds to 20 ms of audio sampled at 16 kHz. The context length of 3,072 tokens, therefore, corresponds to 61.44 s of audio. We refer to these tokens as representations for *frames* of speech.

SpiRit-LM is a continuously pretrained, multimodal variant of Llama-2 (Nguyen et al., 2025; Touvron et al., 2023). The 7B-parameter model was further pretrained on strings of interleaved text and speech tokens, meaning it can take speech as input and generate it. Its speech tokens are also obtained from

discretized features of a HuBERT variant, with each speech token corresponding to 40 ms of speech sampled at 16 kHz.

Llama-Mimi trains Llama-3 (Grattafiori et al., 2024) for speech language modeling using representations from Mimi (Défossez et al., 2024), the same neural codec used by our proposal distribution. While Llama-Mimi extends Llama's vocabulary, it does not interleave text and speech during further pretraining. We use both the 1.3B and 8B variants of Llama-Mimi, which are both trained on 240k hours of speech data. All models are trained and evaluated on English speech.

D GAMM Data Summary

Condition	Coverage	N obs.	Train	Test
Word-level				
GPT-2	–	29,150	26,235	2,915
GSLM (Unmarg.)	–	19,418	17,476	1,942
GSLM (Marg.)	959/1,229 (78.0%)	19,418	17,476	1,942
SpiRit (Unmarg.)	–	10,841	9,757	1,084
SpiRit (Marg.)	809/1,229 (65.8%)	10,841	9,757	1,084
Llama-Mimi (Unmarg.)	–	22,652	20,387	2,265
Llama-Mimi (Marg.)	1,010/1,229 (82.2%)	22,652	20,387	2,265
Llama-Mimi-8B (Unmarg.)	–	22,515	20,264	2,252
Llama-Mimi-8B (Marg.)	1,009/1,229 (82.1%)	22,515	20,264	2,252
Phone-level				
GSLM (Unmarg.)	–	30,763	27,687	3,076
SpiRit (Unmarg.)	–	29,295	26,366	2,930
Llama-Mimi (Unmarg.)	–	29,292	26,363	2,929
Llama-Mimi-8B (Unmarg.)	–	29,292	26,363	2,929
Frame-level				
GSLM (Unmarg.)	–	11,571	10,414	1,157
SpiRit (Unmarg.)	–	22,262	20,036	2,226
Llama-Mimi (Unmarg.)	–	31,582	28,424	3,158
Llama-Mimi-8B (Unmarg.)	–	31,582	28,424	3,158

Table 1: Per-GAMM data summary. *Coverage* is reported only for the marginalized word estimators, and refers to the number of corpus tokens for which the estimator produced an estimate. It is omitted (–) elsewhere, where every token is scored by construction. N obs. is the number of pupil samples entering the model, with the mean per-fold train/test split under 10-fold CV (at $\omega = 1000$ ms). For each speech LM, N obs. also reflects that the word with the largest marginalized surprisal estimate is excluded as an outlier, as the most extreme estimates are most affected by error in importance sampling (see §B). For Llama-Mimi, for example, the largest estimate was 484 nats versus 266 nats at the next-largest value. Marg. and Unmarg. word-level models of a given speech LM are fit on the same folds, so their N and fold sizes are identical.

E Supplementary Results

E.1 GPT-2 Results

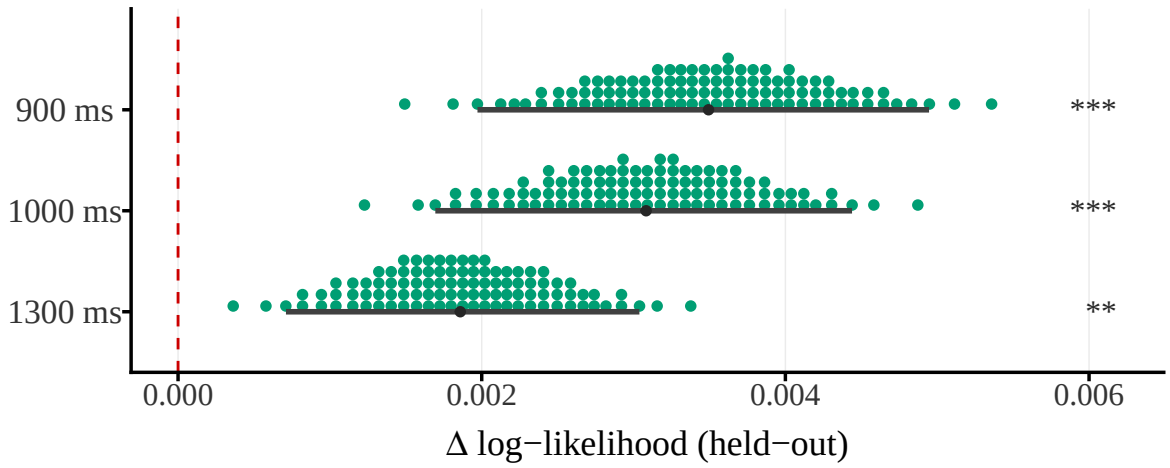

Figure 6: Held-out delta log-likelihood predicting change in pupil size at three offset values ω. Models are fit with word-level surprisal values obtained from GPT-2 using transcriptions of the speech heard by participants. Stars correspond to significant improvements over the no-surprisal baseline; corrected p-values at α levels of 0.05 (*), 0.01 (**), and 0.001 (***).

For reference, we also use GPT-2 (Radford et al., 2019) to calculate surprisal for each orthographic word in the text transcriptions of our stimuli. We denote tokens from GPT-2 as $\phi \in \Phi$ and calculate surprisal as follows:

$$\iota_{\text{gpt}}\left(\phi_t, \boldsymbol{\phi}_{<t}\right) = -\log \overrightarrow{p_{\Phi}}\left(\phi_t \mid \boldsymbol{\phi}_{<t}\right). \tag{23}$$

Because the tokenizer of GPT-2 sometimes splits orthographic words across tokens, we sum the surprisals of any multi-token orthographic words. Consistent with other surprisal theory literature, results indicate that adding GPT-2 surprisals to models of pupil size increases the delta log-likelihood significantly over the baseline model (see Fig. 6).

E.2 Tables of Values

E.3 Marginalized and Unmarginalized Comparisons

E.4 Pupil Time Baseline

Condition	900 ms	1000 ms	1300 ms
GPT-2	0.0035 [0.0020, 0.0049] ***	0.0031 [0.0017, 0.0044] ***	0.0019 [0.0007, 0.0030] **
GSLM Marg.	0.0045 [0.0031, 0.0060] **	0.0041 [0.0028, 0.0054] **	0.0021 [0.0011, 0.0030] **
GSLM Unmarg.	−0.0005 [−0.0016, 0.0006] n.s.	−0.0004 [−0.0014, 0.0006] n.s.	−0.0003 [−0.0010, 0.0004] n.s.
SpiRit Marg.	0.0024 [0.0009, 0.0038] **	0.0018 [0.0006, 0.0030] *	0.0015 [0.0005, 0.0025] **
SpiRit Unmarg.	−0.0034 [−0.0050, −0.0019] n.s.	−0.0031 [−0.0044, −0.0018] n.s.	−0.0029 [−0.0040, −0.0018] n.s.
Llama-Mimi Marg.	0.0021 [0.0007, 0.0035] **	0.0018 [0.0004, 0.0031] *	0.0014 [0.0003, 0.0025] *
Llama-Mimi Unmarg.	0.0006 [−0.0001, 0.0013] n.s.	0.0006 [−0.0001, 0.0012] n.s.	0.0001 [−0.0004, 0.0005] n.s.
Llama-Mimi-8B Marg.	0.0032 [0.0019, 0.0045] **	0.0029 [0.0016, 0.0041] **	0.0024 [0.0014, 0.0035] **
Llama-Mimi-8B Unmarg.	0.0010 [0.0003, 0.0017] *	0.0009 [0.0003, 0.0016] *	0.0006 [0.0001, 0.0010] *

Table 2: Mean and 95% confidence intervals for delta log-likelihood predicting change in pupil size at three offset values ω. Models are fit with word-level surprisal values obtained from four speech LMs and GPT-2. Stars correspond to significant improvements over the no-surprisal baseline; corrected p-values at α levels of 0.05 (*), 0.01 (**), and 0.001 (***).

Condition	900 ms	1000 ms	1300 ms
GSLM Unmarg.	0.0018 [0.0011, 0.0025] **	0.0016 [0.0010, 0.0023] **	0.0011 [0.0007, 0.0016] **
SpiRit Unmarg.	0.0025 [0.0017, 0.0033] **	0.0024 [0.0016, 0.0031] **	0.0017 [0.0010, 0.0023] **
Llama-Mimi Unmarg.	0.0021 [0.0013, 0.0030] **	0.0018 [0.0011, 0.0026] **	0.0012 [0.0006, 0.0018] **
Llama-Mimi-8B Unmarg.	0.0023 [0.0014, 0.0032] **	0.0020 [0.0012, 0.0028] **	0.0012 [0.0006, 0.0018] **

Table 3: Mean and 95% confidence intervals for delta log-likelihood predicting change in pupil size at three offset values ω. Models are fit with phone-level surprisal values obtained from four speech LMs. Stars correspond to significant improvements over the no-surprisal baseline; corrected p-values at α levels of 0.05 (*), 0.01 (**), and 0.001 (***).

Condition	900 ms	1000 ms	1300 ms
GSLM Unmarg.	0.0019 [0.0007, 0.0031] **	0.0009 [0.0001, 0.0017] n.s.	0.0010 [0.0001, 0.0018] n.s.
SpiRit Unmarg.	0.0027 [0.0018, 0.0037] **	0.0027 [0.0017, 0.0036] **	0.0014 [0.0007, 0.0021] **
Llama-Mimi Unmarg.	0.0035 [0.0025, 0.0044] **	0.0030 [0.0021, 0.0039] **	0.0018 [0.0010, 0.0025] **
Llama-Mimi-8B Unmarg.	0.0023 [0.0015, 0.0032] **	0.0020 [0.0011, 0.0028] **	0.0011 [0.0004, 0.0018] *

Table 4: Mean and 95% confidence intervals for delta log-likelihood predicting change in pupil size at three offset values ω. Models are fit with frame-level surprisal values obtained from four speech LMs. Stars correspond to significant improvements over the no-surprisal baseline; corrected p-values at α levels of 0.05 (*), 0.01 (**), and 0.001 (***).

Contrast	900 ms	1000 ms	1300 ms
GSLM	0.0050 [0.0032, 0.0068] $p < 0.001$	0.0045 [0.0029, 0.0061] $p < 0.001$	0.0024 [0.0011, 0.0036] $p < 0.001$
Llama-Mimi	0.0015 [0.0001, 0.0030] $p = 0.019$	0.0012 [−0.0002, 0.0026] $p = 0.044$	0.0014 [0.0003, 0.0025] $p = 0.007$
Llama-Mimi-8B	0.0022 [0.0008, 0.0035] $p = 0.001$	0.0019 [0.0006, 0.0032] $p = 0.002$	0.0019 [0.0008, 0.0029] $p < 0.001$
SpiRit	0.0058 [0.0038, 0.0078] $p < 0.001$	0.0049 [0.0032, 0.0066] $p < 0.001$	0.0044 [0.0031, 0.0058] $p < 0.001$

Table 5: Paired-bootstrap tests for differences between marginalized and unmarginalized word surprisal GAMMs by model and time offset. Values represent the average difference in delta log-likelihood between GAMMs fit with marginalized and unmarginalized word surprisals derived from the same speech LMs. Positive values indicate higher delta log-likelihoods when fit with marginalized surprisals. Confidence intervals are 95% bootstrap intervals.

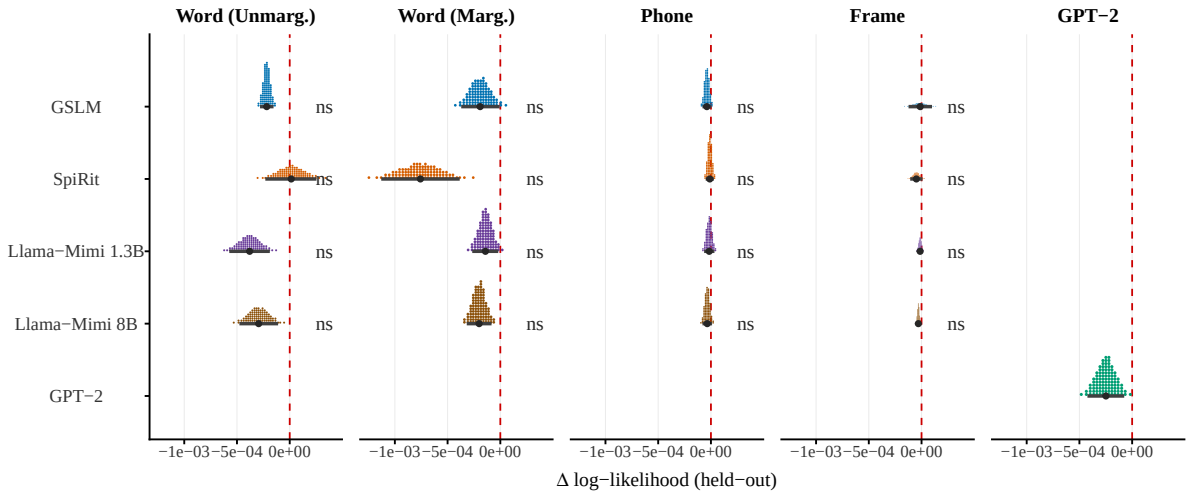

Figure 7: Held-out delta log-likelihood relative to baseline model of pupil size, augmented with a smooth for the time in the trial at which the pupil measurement was taken (offset: $\omega = 1000$ ms). Each dot represents 1% of the bootstrap distribution of the mean delta log-likelihood; the point and bars below give the bootstrap mean and 95% CI.