GRAMMAR-WEIGHTED FUTURES: MODELING HUMAN EXPECTATION WITH OPEN GOALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Behavioral and neural evidence suggests that human expectations in language and music are not driven by surface statistics alone: readers and listeners also anticipate how open goals will be completed, as a subject awaits its verb and a chord progression awaits its return to the home chord. We present Grammar-Weighted Futures, a model in which such anticipation constrains the probability of the next event. A fixed surface sequence model, which predicts events well without exposing open goals, samples continuations; a probabilistic grammar, which represents open goals explicitly but predicts events poorly, scores how well each continuation completes the goals of the observed prefix; and these scores reweight the surface prediction of the next event. Tested against listeners' surprise ratings for chord progressions with Transformer, LSTM and variable-order surface models, our proposed grammar weighting improves the prediction of held-out ratings with all three. In language, an exploratory implementation improves the prediction of word-predictability ratings over GPT-2 alone. These results indicate that expectations about how open goals will be completed shape human expectation in abstract cognitive domains such as language and music beyond surface statistics.

1 INTRODUCTION

Theories of expectation in language and music commonly quantify how expected an event is by its probability given the preceding context (Hale, 2001; Levy, 2008; Pearce & Wiggins, 2012). Yet readers and listeners also anticipate how structure that has begun will be completed. Readers who encounter *either* anticipate a later *or* (Staub & Clifton, 2006); holding predicted syntactic heads in memory incurs processing cost (Gibson, 1998; Chen et al., 2005); and cortical activity tracks the number of open constituents and drops when they close (Nelson et al., 2017). Music has comparable structure. A chord is several notes sounded together, and a chord progression is a sequence of chords; each piece has a key whose home chord (the *tonic*, written I) sounds stable, so a progression sets up expectations about returning to it. Listeners' brain responses distinguish progressions that return to their initial key from those that do not (Koelsch et al., 2013), and hierarchical models predict listeners' judgments of how a progression will be completed (Herff et al., 2021). We call a constituent that the context predicts but has not yet begun an *open goal*; once begun, it remains open until it is completed. An open goal need not determine the next event: a progression may linger on other chords before it returns home. Maintaining it therefore need not improve next-event prediction, the criterion by which computational models of expectation are usually trained and evaluated. We ask whether an explicit account of open goals and their possible completions nonetheless improves the modeling of human expectation. Our hypothesis is that expectations about how open goals will be completed constrain the probabilities of upcoming events, even where they do not raise the probability of what occurs.

Two modeling traditions address distinct parts of this question. Surface sequence models, from variable-order Markov models such as IDyOM to neural networks, predict the next event well; their predictions can reflect internally encoded structure, but they do not expose whether an open goal mediates a prediction (Pearce & Wiggins, 2006; 2012). Probabilistic grammars represent open goals explicitly (Lerdahl & Jackendoff, 1983; Rohrmeier, 2011), and prefix parsers derive next-event probabilities by marginalizing grammatical continuations (Hale, 2001; Levy, 2008; Stolcke, 1995), but grammars alone are weak sequence predictors. Rather than comparing the two (Frank & Bod, 2011; Fossum & Levy, 2012), we combine them: the surface model supplies fluent continuations, and the grammar supplies the criterion by which they are weighed.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

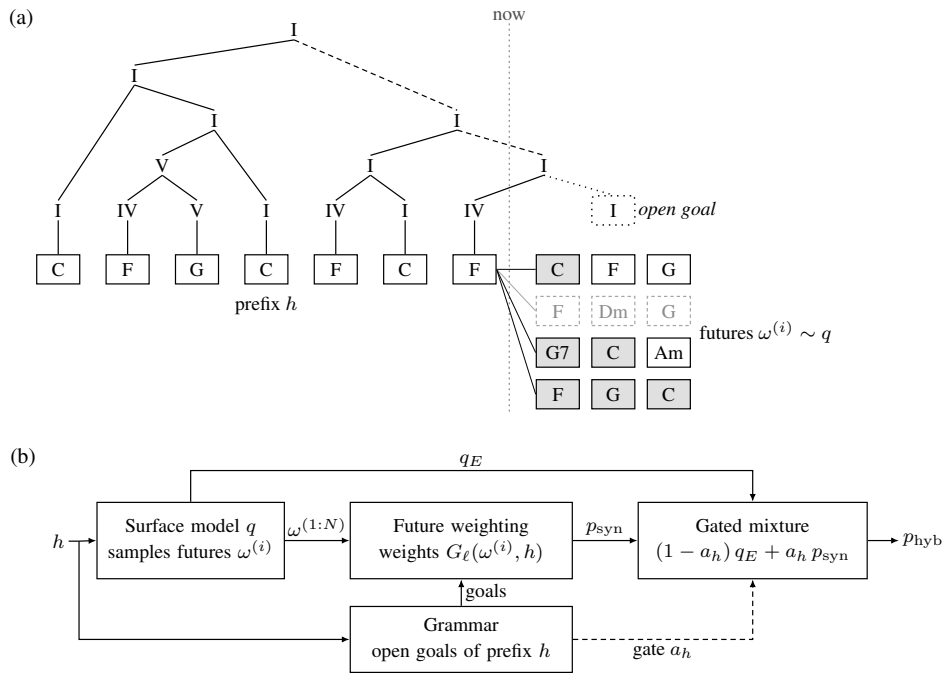


Figure 1: **Grammar-weighted futures.** (a) Illustrative schematic. Roman numerals mark each chord's role in the key; I is the home chord, C. Solid branches mark completed constituents, dashed ones constituents still growing, and the dotted node is an open goal, a predicted return to I. Sampled futures that reach C complete it (shaded) and score high; the grey one does not. (b) Computation. The surface model samples futures of h; the grammar infers its open goals. Each future is scored by how well it completes them (G_ℓ, at each completion length ℓ). Adding up the normalized scores of futures that begin with the same event gives the grammar-weighted next-event distribution p_{syn}. A gate a_h computed from the prefix analysis (dashed) mixes p_{syn} with the surface model's unchanged prediction q_E (top line). The mixture's surprisal, $-\log_2 p_{\mathrm{hyb}}$, predicts human responses in a separate model; responses never enter this computation. Appendix D.1 gives a computed example.

Our model, *Grammar-Weighted Futures*, makes this division of labor explicit (Figure 1): a fixed surface model samples possible futures, a probabilistic grammar scores how well each future completes the open goals of the prefix, and the scores are turned into a next-event distribution that a gate mixes with the surface model's own prediction. If every future receives the same score, the model returns the surface prediction. The hypothesis is stated at the computational level: we do not assume that people draw Monte Carlo samples.

A gain from grammar weighting would not by itself show that open goals matter, so we test three claims. *Representation* (H1): weighting futures by the grammar improves the prediction of human expectation over the surface model alone; in music we also compare it with a static preference for stable chords (Krumhansl & Kessler, 1982; Lerdahl, 2001). *Alignment* (H2): the gain comes from the grammar's evaluation of each future rather than from reweighting as such, so shuffling the scores among the futures or sharpening the surface prediction to the same entropy should not reproduce it. *Prospection* (H3): the gain comes from evaluating how futures continue beyond the next event, so combining the surface prediction only with the grammar's own prediction of the next event should fall short.

We test these claims in two higher-level cognitive domains: music and language. In music, where chords are the grammar's own terminal symbols and listeners rated their surprise chord by chord (Cheung et al., 2024), we tested Transformer, LSTM and IDyOM surface models with a grammar that describes each chord by its role in the current key (Harasim et al., 2018). An exploratory language implementation that combines GPT-2 (Radford et al., 2019) with a treebank grammar improves the prediction of word-predictability ratings (de Varda et al., 2024) over GPT-2 alone, with an uncertain increment beyond incremental alternative sampling (IAS; Giulianelli et al., 2026).

Our contributions are twofold:

1. *Grammar-Weighted Futures*, a way to add a probabilistic grammar's knowledge of open goals to a trained sequence model without retraining it. The grammar reweights the model's sampled continuations by how well they complete open goals.
2. Evidence that this knowledge improves models of human expectation without improving event prediction. Grammar weighting improves the prediction of listeners' surprise ratings for three sequence models without making the chords that occur more probable.

2 Related Work

Surprisal and human expectation. Surprisal, or information content (IC), is the standard link between predictive models and human expectation in music (Pearce & Wiggins, 2012) and in language, where it predicts reading times across eleven languages (Wilcox et al., 2023) and N400 amplitudes (Michaelov et al., 2024). Better next-event prediction does not guarantee a better account of behavior. Psychometric fit first improved with language-model quality (Goodkind & Bicknell, 2018), but not in Japanese (Kuribayashi et al., 2021) and not for larger pretrained Transformers (Oh & Schuler, 2023). These findings compare different models. We hold the surface model and its sampled futures fixed and change only how probability is distributed among its predictions.

Statistical and structural expectation. In music, statistical models such as IDyOM learn sequential regularities (Pearce & Wiggins, 2006) and predict listeners' surprise at chords (Cheung et al., 2024). Schematic knowledge of keys offers another route (Huron, 2006): listeners rank the notes of a key by how stable they sound (Krumhansl & Kessler, 1982) and process chords closely related to preceding chords faster (Bharucha & Stoeckig, 1986; Bharucha, 1987; Tillmann et al., 2000), neither of which requires an open goal. Generative grammars of music make nested dependencies between chords explicit (Rohrmeier, 2011; Harasim et al., 2018), hierarchical models of tonal tension fit listeners' tension ratings (Lerdahl, 2001; Lerdahl & Krumhansl, 2007), and the number of open dependencies in such analyses predicts theta-band MEG activity (Herff et al., 2024). In language, whether hierarchical models explain reading beyond sequential ones depends on lexical controls (Frank & Bod, 2011; Fossum & Levy, 2012), recurrent neural network grammars fit EEG and reading times better than LSTMs (Hale et al., 2018; Yoshida et al., 2021), and incremental parsers separate lexical from syntactic surprisal (Roark et al., 2009). Open goals match predicted categories of Gibson (1998), predicted nodes of Demberg et al. (2013) and open implications of Rohrmeier (2013).

Sampled futures and reweighting. Anticipation has been measured as a grammar's uncertainty about the rest of the sentence (Hale, 2006; Linzen & Jaeger, 2016), as contextual entropy (Pimentel et al., 2023) and, more generally, as expectations of functions over sampled continuations (Giulianelli et al., 2024). Our model belongs to this family, with a continuation function given by an external grammar potential and normalized into a next-event distribution. IAS, another member, measures target-conditioned change in language-model representations of sampled alternatives (Giulianelli et al., 2026), and we evaluate it as a close alternative. Reweighting sampled futures by a potential is established in controlled generation (Yang & Klein, 2021; Zhao et al., 2024; Loula et al., 2025). Here the potential comes from goals inferred from the prefix, and the target is human expectation.

3 Method

The grammar-weighted futures model implements the hypothesis that expectations about how open goals will be completed constrain next-event probabilities. Given an observed prefix h, the chords or words heard so far, a fixed surface model assigns a probability $q_E(c \mid h)$ to each possible next event c. Every probability model in this paper turns this surface prediction into a hybrid prediction of the same form:

$$p_{\text{hyb}}(c \mid h) = q_E(c \mid h)\big[(1 - a_h) + a_h \Psi_h(c)\big], \qquad \mathbb{E}_{q_E}[\Psi_h] = 1. \tag{1}$$

Here $\Psi_h(c) \geq 0$ is an event-specific correction factor: events with $\Psi_h(c) > 1$ become more probable and events with $\Psi_h(c) < 1$ less probable. The gate $a_h \in [0, 1]$, computed from the prefix, sets how

strongly the correction is applied. Because Ψ_h averages to one under q_E, p_{hyb} remains a probability distribution, and $\Psi_h \equiv 1$ returns the surface prediction unchanged.

Most comparisons in Section 4 keep q_E and a_h fixed and change only how Ψ_h is obtained. They are therefore comparisons of correction factors: two models with the same Ψ_h assign the same probability to every event the surface model considers possible, however Ψ_h was computed. The prospection test (H3) accordingly asks whether scoring how the surface model's futures complete open goals yields a different Ψ_h from the grammar's own prediction of the next event, and whether the difference predicts behavior. We describe the full model for music, the primary test. It obtains Ψ_h in three steps (Figure 1): it samples futures from the surface model (§3.1), infers the grammar's open goals (§3.2), and scores how the futures complete them (§3.3). Sections 3.4 and 3.5 define the gate and the language version.

3.1 Step 1: sampling futures from a surface model

Human responses are recorded for *events*: chords in music, and complete words or punctuation marks in language. The surface model q is autoregressive: it predicts one token at a time from the tokens before it, so it can also sample continuations ω of h, which we call futures. Futures are possibilities, not observed events. We write $\pi(\omega)$ for the first event of ω, so that formally $q_E(c \mid h) = \Pr_q[\pi(\omega) = c \mid h]$.

In music, a Transformer, an LSTM and IDyOM, a variable-order Markov model that mixes corpus and current-piece statistics (Pearce & Wiggins, 2006), share a vocabulary of 739 chord symbols. Each symbol gives the chord's root (the note it is built on) relative to the song's annotated key, its quality (such as major, minor or seventh) and its inversion (which note is lowest), and each chord is one token, so q_E is simply the model's next-chord distribution. The song's overall key is given to the models; brief moves to other keys within the song are left to the grammar. Apart from this key, everything the model uses is computed from the prefix.

3.2 Step 2: the grammar's open goals

A probabilistic grammar is a set of rewrite rules with probabilities that generates sequences as trees. Both of our grammars are estimated from corpora, never from human responses (Appendix B). The chord grammar, adapted from Harasim et al. (2018), has 168 categories, the seven scale degrees (a chord's position within a key, written I–VII) in each of 24 major and minor keys. Its rules cover prolongation (repeating a chord's role), preparation (a chord that leads into another, such as V before I), tonicization (briefly treating another degree as a local home chord) and substitution (replacing a chord by another with the same role). Its rule probabilities are learned from the McGill Billboard chord corpus (Burgoyne et al., 2011) without annotated trees.

Given h, an incremental parser finds the partial trees (analyses) that are consistent with the prefix and computes their probabilities (Jelinek & Lafferty, 1991; Stolcke, 1995). In each analysis, the rules already used predict constituents that have not yet begun; their entire yields, the terminals they generate, lie in the future. These are the prefix's *open goals*. We track the first one or two open goals, the *tracked goals*, and sum out later ones. This gives posterior probabilities $p_1(g_1 \mid h)$ for analyses with one open goal and $p_2(g_1, g_2 \mid h)$ for ordered pairs of goals; F denotes the first goal. Two goals also capture what must follow the first: in language, after *Alice*, a toy grammar leaves a verb phrase and then sentence-final punctuation, $(\mathrm{VP}, \mathrm{STOP})$, open (Appendix B.2). The same probabilities give $d_h(\ell)$, the grammar's probability that completing the tracked goals takes ℓ terminals, truncated at a horizon H and renormalized over $1 \le \ell \le H$ (Eq. 10 in Appendix B.2).

3.3 Step 3: from scored futures to the correction

At each length $\ell \le H$, a sampled future receives a nonnegative score $G_\ell(\omega, h)$ for how its first ℓ terminals complete the tracked goals. In music, G_ℓ is the probability that the first ℓ chords are exactly the yield of the tracked goals:

$$G_\ell^{\mathrm{music}}(\omega, h) = \sum_{g_1} p_1(g_1 \mid h)\, \beta_{g_1}(\omega_{1:\ell}) + \sum_{k=1}^{\ell-1} \sum_{g_1, g_2} p_2(g_1, g_2 \mid h)\, \beta_{g_1}(\omega_{1:k})\, \beta_{g_2}(\omega_{k+1:\ell}), \qquad (2)$$

where $\beta_g(\cdot)$, the inside probability, is the probability that category g generates exactly the given chords (reduced to root and one of ten chord qualities), and k ranges over all ways of splitting the ℓ chords between the two goals. G_ℓ thus includes the grammar's learned probabilities of chord qualities.

Scores are then normalized within each length, so that they average to one over futures, and averaged over lengths with the weights $d_h(\ell)$. The result, $W_h(\omega)$, is the future's weight, and the correction factor for an event is the average weight of the futures that begin with it:

$$\Psi_h(c) = \mathbb{E}_q\big[W_h(\omega) \mid \pi(\omega) = c, h\big], \qquad W_h(\omega) = \sum_{\ell=1}^{H} d_h(\ell)\, \frac{G_\ell(\omega, h)}{Z_\ell(h)}, \tag{3}$$

with $Z_\ell(h) = \mathbb{E}_q[G_\ell(\omega, h)]$, so that $\mathbb{E}_q[W_h] = 1$ as Eq. 1 requires. Events that begin well-completing futures are raised, and events whose futures leave the goals unmet are lowered. The resulting grammar-weighted next-event distribution is $p_{\text{syn}} = q_E \Psi_h$. We estimate it from N sampled futures by self-normalized importance sampling: for each length, $p_{\text{syn}}(c \mid h)$ takes the score-weighted fraction of futures that begin with c, and mixes these fractions over lengths with weights $d_h(\ell)$ (Appendix A, Eq. 6; sample sizes in Appendix B.3).

3.4 Step 4: the gate

The gate a_h in Eq. 1 is computed from the grammar's analysis of the prefix, not fitted to responses. Equivalently, $p_{\text{hyb}} = (1 - a_h)\, q_E + a_h\, p_{\text{syn}}$, a mixture of the surface and grammar-weighted predictions. In music, the gate reflects how certain the prefix analysis is about the first goal F: $a_h = 1 - \mathbb{H}[p(F \mid h)]/\log_2 168$, where $\mathbb{H}$ is entropy in bits, so that a_h is one when the analysis is certain of F and zero when all 168 categories are equally likely. Both q_E and p_{hyb} receive the same uniform smoothing, which spreads 3.4% of the mass over all chords (Appendix B.3).

3.5 Language instantiation

Language uses the same construction with GPT-2 small (Radford et al., 2019) as the surface model and a probabilistic context-free grammar (PCFG) estimated from the parsed Penn Treebank (PTB) sample distributed with NLTK (Marcus et al., 1993), as in PCFG models of syntactic surprisal (Hale, 2001; Levy, 2008). Three components change; Table 3 in Appendix B summarizes both instantiations.

Events. GPT-2 predicts byte-pair-encoding (BPE) tokens and the grammar generates PTB terminals, neither of which always coincides with a word: one word may span several BPE tokens, and *don't* corresponds to two PTB terminals. A lexer reads the next complete word or punctuation mark from each sampled continuation, and because a word can be spelled by several token sequences, q_E sums over all of them (Pimentel & Meister, 2024); Appendix A gives this estimator (Eq. 4).

Scores. Sentences rarely complete their goals within a few words, so a completion probability would give most futures zero weight; the language score instead credits progress. Future words enter only through their part-of-speech (POS) tags: a word-to-tag model trained on the same treebank gives each word a distribution over tags, and the grammar's own word probabilities are summed out because GPT-2 already models word frequency. After each future terminal, the parser divides the probability of the analyses still consistent with the future, $m_s(\omega, h, \ell)$, among four states s: the first goal is still open, the first of two goals is complete, the only goal is complete, or both are complete. The score is their utility-weighted average, $G_\ell^{\text{lang}}(\omega, h) = \sum_s u_s m_s(\omega, h, \ell) / \sum_s m_s(\omega, h, \ell)$ with $u = (1, 2, 2, 3)$, that is, one plus the expected number of goals completed. A future that no analysis can continue scores zero (Appendix B.2).

Gate. Phrase-level goals constrain the next word mainly through its syntactic category, so the gate is the fraction of uncertainty about the next POS terminal T that the first goal explains, $a_h = \mathbf{1}\{m_h \geq 4\}\, I(F; T \mid h)/\mathbb{H}[p(T \mid h)]$, where I is mutual information and m_h the number of PTB terminals in the prefix. The gate is zero when T is certain or when the prefix has fewer than four terminals, because analyses of the first few words are unreliable (Appendix B.2).

4 EXPERIMENTS

4.1 SETUP

Musical surprise is the primary test of H1–H3, and language gives an exploratory test of H1. We measure how surprising each event is to a model by the information content (IC, $-\log_2 p(x \mid h)$, in bits) of the event x that occurred. The *Baseline* condition takes the IC from the surface model's own prediction q_E, and *Ours* takes it from the grammar-weighted prediction p_{hyb} of our model (Eq. 1); each control changes one part of our model (Table 1). The IC then enters a separate response model of the human ratings. No part of the probability models is fitted to human responses (Appendix C).

How results are reported. Each comparison asks which of two conditions better predicts human ratings that were held out from the response model. The measure is the expected log predictive density of the held-out ratings (ELPD, in nats), summed over all ratings, and a comparison reports the difference, ΔELPD; a positive value favors the condition named first. Two further numbers convey its size. Dividing ΔELPD by the number of ratings gives the average gain per rating: in music, a difference of $+20$ over the 2,018 ratings means that each held-out rating receives on average (geometric mean) about 1% more predictive density. We also count the songs in which the first condition predicts better. Uncertainty is a 95% interval from a paired bootstrap over the 30 songs (over sentences in language), and we call a difference *reliable* when this interval excludes zero. Standard errors and the remaining intervals are given in the appendix tables.

Music. The surprise data of Cheung et al. (2024) contain chord-by-chord surprise ratings of 30 chord progressions by 25 participants. We predict the mean rating of the musicians and of the nonmusicians at each of the 1,009 chords after the first, 2,018 ratings in all. The Transformer, LSTM and IDyOM are trained in ten song-level folds of the 739-song McGill Billboard corpus (Burgoyne et al., 2011), and each stimulus is scored by the models whose fold excludes its source song; the chord grammar is estimated once, from the 709 songs that are not stimulus sources (Appendix B.1). All three surface models share the grammar, $H = 10$, gate and smoothing. Because the futures are sampled, every music predictor is computed from three independent sets of samples (streams), each fitted separately, and the pointwise results are averaged over them. The main analysis uses the response model released with the data and predicts each held-out rating from the other ratings in its song (conditional response-level leave-one-out, response-LOO). A second analysis holds out whole songs and predicts all their ratings from the other 29 songs (whole-song holdout; Appendix C). Results refer to the main analysis unless whole songs are said to be held out.

Language. The resource of de Varda et al. (2024) provides predictability ratings for 1,726 word events in 205 English sentences. We reverse the ratings, so that larger values align with greater IC, and predict them in ten folds that each hold out whole sentences. Ours combines released GPT-2 small (Radford et al., 2019) with a PCFG and word-to-tag model estimated from the separate 3,914-sentence Penn Treebank sample distributed with NLTK (Marcus et al., 1993), using $H = 2$ PTB terminals. Because this configuration was developed on the same resource, the result is exploratory.

Transfer. Secondary analyses test transfer to continuous EEG recorded while listeners heard Bach chorales (Teng et al., 2024) and to reading times, N400 and cloze responses (Appendix D.3).

4.2 GRAMMAR WEIGHTING AND MUSICAL SURPRISE (H1)

Does weighting futures by the grammar improve the prediction of listeners' surprise? Yes, for all three surface models (Figure 2a). Ours predicted the held-out ratings better than Baseline in 23 of the 30 songs for the LSTM, in 21 for the Transformer and in 27 for IDyOM. Per rating, Ours assigns on average 3.1%, 2.4% and 5.4% more predictive density. All three gains are reliable, also when whole songs are held out (Appendix D.1).

Does the gain come from predicting the chords that occurred better? No (Figure 2b). For the two neural surface models, Ours made the chords that actually occurred less probable, by 10.4% and 9.5% on average (geometric mean), and it did so in 23 and 25 of the 30 songs. For IDyOM, the change is not reliable. Ours thus predicts listeners' ratings better without predicting the chords themselves better.

Table 1: **Conditions compared in music.** Each condition forms its next-chord prediction by changing the surface model's prediction q_E, and all share the same smoothing. Comparing a control with Ours answers the question in the last column.

Condition	How the next-chord prediction is formed	Question answered by the comparison with Ours
Baseline	The surface model alone, q_E	Does grammar weighting help at all? (H1)
Ours	q_E reweighted by how well its sampled futures complete the open goals, with the gate (Section 3)	—
Permuted	As Ours, but the scores are shuffled among the same sampled futures	Is the gain due to the grammar's evaluation of each future, or only to reweighting? (H2)
Entropy-matched	q_E made exactly as sharp or as flat as our model's prediction, without grammar scores	Is the gain only a sharper or flatter prediction? (H2)
One-step	q_E multiplied by the grammar's own probability of the next chord, with our model's gate	Do the surface model's futures add to the grammar's next-chord prediction? (H3)
Static tilt	q_E multiplied by a fixed preference over chord categories, with our model's gate	Is a context-independent preference for stable chords enough?

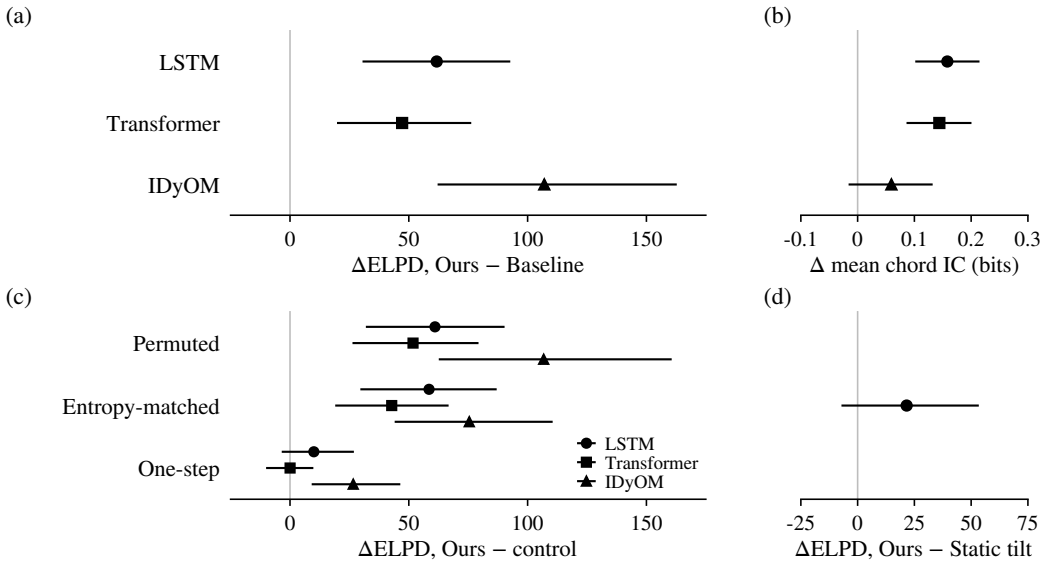


Figure 2: **Musical surprise.** Points in (a), (c) and (d) show how much better Ours predicts held-out ratings than another condition (ΔELPD; positive values favor Ours). Whiskers are 95% paired bootstrap intervals over the 30 songs, after averaging the three sampling streams. (a) Ours compared with Baseline, for each surface model. (b) Change in the mean IC of the chords that occurred, Ours minus Baseline, in bits; positive values mean that Ours made these chords less probable. (c) Ours compared with each control (circle, LSTM; square, Transformer; triangle, IDyOM); the LSTM is the primary test. (d) Ours compared with Static tilt for the LSTM, with whole songs held out.

4.3 What produces the gain (H2, H3 and a static tonal prior)

The next tests ask which part of our model produces the gain. Each control changes one part of it (Table 1) while reusing the same sampled futures, response model and smoothing. The LSTM is the primary test, with decision rules fixed before the results were computed; the Transformer and IDyOM repeat the H2 and H3 comparisons as consistency analyses (Figure 2c; Appendix D.1).

H2: does the gain come from the grammar's evaluation of each future, or just from reweighting? From the grammar's evaluation. Any reweighting changes the surface prediction, for instance by making it sharper, and such a change alone might improve the fit. Permuted reweights the same futures with the same set of grammar scores, but shuffled among them, so the scores no longer describe the futures they are attached to. Entropy-matched makes the surface prediction as sharp or as flat as our model's without using the grammar. Neither control improves reliably on Baseline for the LSTM, and Ours is reliably better than both controls for all three surface models (Figure 2c). The one qualification is IDyOM, for which sharpening alone recovers part of the gain (Appendix D.1).

H3: do the surface model's futures add to the grammar's own next-chord prediction? Not detectably for the LSTM. The grammar can predict the next chord by itself: its probability of the next chord already averages over the grammar's own ways of completing the open goals, but it ignores the surface model's futures. One-step multiplies the surface prediction by this probability, with our model's grammar, gate and smoothing. The decision rule was as follows. If One-step improves on Baseline and Ours is not reliably better than One-step, the gain is available from the grammar's next-chord prediction; if Ours is reliably better, part of the gain is specific to scoring the futures.

For the LSTM, One-step alone recovers most of the gain ($+51.7$ of $+61.7$). Ours is better than One-step in only 16 of the 30 songs, and the difference is not reliable ($+10.0$, $[-3.5, 26.9]$). Under the decision rule, most of the LSTM's gain is therefore available from the grammar's next-chord prediction. The answer depends on the surface model (Figure 2c): Ours is reliably better than One-step for IDyOM and essentially identical to it for the Transformer. One-step also differs from Ours restricted to one-chord futures ($H = 1$), which credits a future only when its first chord completes a goal by itself and barely improves on Baseline; with two-chord futures, Ours already reaches its full gain for the neural models (Appendix D.1).

4.4 Language (exploratory)

Does grammar weighting also help in language? Over GPT-2 alone, yes. Ours predicted the reversed predictability ratings reliably better than Baseline, and did so in 114 of the 205 sentences (Table 2). Four simpler syntax measures, each entered as an additional predictor beside Baseline, do not account for this gain (Appendix D.2). The gain peaks at futures of $H = 2$ PTB terminals; because some rated words span several terminals, this is a sensitivity result, not evidence that readers anticipate several words ahead.

Does it add to the closest future-based alternative? Not detectably. IAS (Giulianelli et al., 2026) measures how a word changes a language model's representations of sampled continuations. Its released score improves prediction more than Ours does, and adding Ours on top of it gives a positive but uncertain further gain (Table 2; nested-selection variants in Appendix D.2). The result thus supports grammar weighting over GPT-2 alone, not over IAS.

Table 2: Exploratory language results: how much better the first condition predicts held-out ratings (ΔELPD with sentence-cluster SE, and 95% paired bootstrap interval over the 205 sentences). Each of the ten folds holds out whole sentences.

Contrast	ΔELPD $\pm$ SE	95% interval
Ours $-$ Baseline	$+5.080 \pm 2.032$	$[1.188, 9.088]$
(Baseline + released IAS) $-$ Baseline	$+9.574 \pm 4.509$	$[0.629, 18.328]$
Ours - (Baseline + released IAS)	$+3.671 \pm 4.910$	$[-5.981, 13.144]$

4.5 Transfer beyond explicit ratings

Does the gain carry over to other measures of expectation? Not reliably. In the EEG data, chord IC from Baseline or Ours did not improve held-out prediction beyond acoustic features for either surface model, and Ours did not differ from Baseline. In language, reading times and cloze show no reliable gain, and the N400 gain depends on the analysis settings (Appendix D.3).

5 DISCUSSION

We asked whether an explicit account of open goals improves the modeling of human expectation, even when it does not make the events that occur more probable. In music, the answer was yes. We weighted a sequence model's sampled continuations by how well they complete the progression's open goals, such as a return to the home chord. This made listeners' surprise ratings more predictable for all three models, including for songs the response model had never seen.

The gain did not come from predicting the chords better. For the neural models, our model made the chords that occurred about 10% less probable, yet it predicted the ratings better. This dissociation is possible because an open goal need not determine the next event. Similar dissociations are known in language, where models that predict words better need not fit reading behavior better (Kuribayashi et al., 2021; Oh & Schuler, 2023), and language-model surprisal underestimates syntactic processing difficulty (van Schijndel & Linzen, 2021; Arehalli et al., 2022; Huang et al., 2024). Unlike in those studies, the surface model here was held fixed, so the dissociation comes from the grammar's correction alone. The gain also did not come from reweighting. Shuffling the grammar's scores among the continuations, or merely sharpening the surface prediction, did not reproduce it (H2). Instead, part of what the grammar supplies is knowledge of which chords are stable in a key (Krumhansl & Kessler, 1982; Lerdahl, 2001): a fixed preference for stable chords recovered more than half of the LSTM's gain when whole songs were held out. How much the goals of each particular progression add beyond this depended on how the preference was defined.

At first glance, the answer to our secondary question (H3) may seem to undercut the main result. For the neural models, most of the gain remained when the surface model's continuations were not used (One-step). However, One-step still uses open goals. Its next-chord probability comes from a prefix parser, which averages over all the ways the grammar can complete the open goals (Hale, 2001; Levy, 2008; Chelba & Jelinek, 2000; Roark, 2001). Both conditions therefore evaluate the next chord against the open goals; they differ only in whether the continuations come from the surface model or from the grammar. The surface model's continuations added a reliable gain only for IDyOM, which has no learned hierarchical state. Transformers, in contrast, can learn hierarchical structure and anticipate upcoming tokens (Hewitt & Manning, 2019; Manning et al., 2020; Murty et al., 2023; Yao et al., 2021; Pal et al., 2023), which may leave less for their continuations to add. H3 thus refines how the expectation is best computed. Our results show that open goals shape surprise, not that these goals must be deeply nested.

These findings have clear bounds. The main evidence comes from a single study of 30 chord progressions from popular songs, rated chord by chord, with each song's key given to the models. The decision rule for H3 and the comparison with a static tonal prior used the LSTM alone. The language result is exploratory, with a grammar that sees future words only through their part-of-speech tags, and the gain did not transfer reliably to EEG or reading times (Section 4.5). Our claim therefore concerns explicit judgments of surprise.

Taken together, listeners' surprise appears to reflect expectations about how open goals will be completed, beyond the statistics a sequence model learns. For machine learning, the results suggest that next-event likelihood is an incomplete criterion for models of human expectation.

6 CONCLUSION

This paper addressed a long-standing question: does expectation in language and music rest only on the probability of the next event, or also on structure that the context has left open? Surprisal theory and statistical learning have made next-event probability the common currency of expectation (Hale, 2001; Levy, 2008; Pearce & Wiggins, 2012), whereas structural theories hold that readers and listeners also track open goals (Gibson, 1998; Lerdahl & Jackendoff, 1983; Rohrmeier, 2011). Direct comparisons between sequential and hierarchical models have been hard to interpret, because the models differ in more than their structure (Frank & Bod, 2011; Fossum & Levy, 2012). *Grammar-Weighted Futures* instead add a grammar's open goals to a trained sequence model, without retraining it, and measure what they contribute while the sequence model stays fixed. For three such models, the open goals made listeners' surprise ratings more predictable. The relation between learned statistics and structural expectation thus becomes a question that can be tested with other grammars, models and data.

AI USAGE DISCLOSURE

We used Claude Code and OpenAI Codex to assist with conceptual and hypothesis refinement, mathematical exposition and checks, experimental design and parameter suggestions, method implementation, debugging, data processing, and data analysis. These tools also supported literature discovery, figure preparation, manuscript editing, and pre-submission critique. The authors determined the research questions, final experimental design, and scientific claims. They personally reviewed the code and tests, checked mathematical derivations and numerical results, verified cited references against original sources, and reviewed the interpretations and final manuscript against the evidence. AI-generated material was not used as human behavioral or neural observations. The authors take full responsibility for the content of this work, including its text, claims, analyses, and artifacts.

ETHICS STATEMENT

This study reanalyzes the cited behavioral, EEG and corpus resources; it collects no new participant data. Human-subject recruitment and consent are described in the original studies. We do not claim a new ethics approval or permission to redistribute the original recordings and participant data.

REPRODUCIBILITY STATEMENT

Section 3 and Appendices A–C specify the estimator, the grammars, the surface models, the response models and the evaluation units. The supplementary material contains code for every reported experiment, organized by section, with download instructions for all datasets; no third-party data are redistributed. It also contains the estimated chord grammar, the per-event predictors and response-LOO scores behind the reported results, and an artifact map with SHA-256 checksums (Appendix E). From these files, the reported rating and language results can be recomputed on a CPU without retraining the grammar or the surface models; the EEG analysis requires the original recordings. Seeds, software versions and hardware requirements are listed in the accompanying README.

REFERENCES

Suhas Arehalli, Brian Dillon, and Tal Linzen. Syntactic surprisal from neural models predicts, but underestimates, human processing difficulty from syntactic ambiguities. In *Proceedings of the 26th Conference on Computational Natural Language Learning*, pp. 301–313. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.conll-1.20.

Jamshed J. Bharucha. Music cognition and perceptual facilitation: A connectionist framework. *Music Perception*, 5(1):1–30, 1987. doi: 10.2307/40285384.

Jamshed J. Bharucha and Keiko Stoeckig. Reaction time and musical expectancy: Priming of chords. *Journal of Experimental Psychology: Human Perception and Performance*, 12(4):403–410, 1986. doi: 10.1037/0096-1523.12.4.403.

John Ashley Burgoyne, Jonathan Wild, and Ichiro Fujinaga. An expert ground truth set for audio chord recognition and music analysis. In *Proceedings of the 12th International Society for Music Information Retrieval Conference*, pp. 633–638, 2011. doi: 10.5281/zenodo.1417547.

Paul-Christian Bürkner, Jonah Gabry, and Aki Vehtari. Efficient leave-one-out cross-validation for Bayesian non-factorized normal and Student-t models. *Computational Statistics*, 36(2): 1243–1261, 2021. doi: 10.1007/s00180-020-01045-4. URL https://doi.org/10.1007/s00180-020-01045-4.

Ciprian Chelba and Frederick Jelinek. Structured language modeling. *Computer Speech & Language*, 14(4):283–332, 2000. doi: 10.1006/csla.2000.0147.

Evan Chen, Edward Gibson, and Florian Wolf. Online syntactic storage costs in sentence comprehension. *Journal of Memory and Language*, 52(1):144–169, 2005. doi: 10.1016/j.jml.2004.10.001. URL https://doi.org/10.1016/j.jml.2004.10.001.

Vincent K. M. Cheung, Peter M. C. Harrison, Stefan Koelsch, Marcus T. Pearce, Angela D. Friederici, and Lars Meyer. Cognitive and sensory expectations independently shape musical expectancy and pleasure. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 379(1895): 20220420, 2024. doi: 10.1098/rstb.2022.0420.

Andrea Gregor de Varda, Marco Marelli, and Simona Amenta. Cloze probability, predictability ratings, and computational estimates for 205 English sentences, aligned with existing EEG and reading time data. *Behavior Research Methods*, 56(5):5190–5213, 2024. doi: 10.3758/s13428-023-02261-8.

Vera Demberg, Frank Keller, and Alexander Koller. Incremental, predictive parsing with psycholinguistically motivated tree-adjoining grammar. *Computational Linguistics*, 39(4):1025–1066, 2013. doi: 10.1162/COLI_a_00160. URL https://aclanthology.org/J13-4008/.

Victoria Fossum and Roger Levy. Sequential vs. hierarchical syntactic models of human incremental sentence processing. In *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2012)*, pp. 61–69. Association for Computational Linguistics, 2012. URL https://aclanthology.org/W12-1706/.

Stefan L. Frank and Rens Bod. Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, 22(6):829–834, 2011. doi: 10.1177/0956797611409589. URL https://doi.org/10.1177/0956797611409589.

Edward Gibson. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1): 1–76, 1998. doi: 10.1016/S0010-0277(98)00034-1. URL https://doi.org/10.1016/S0010-0277(98)00034-1.

Mario Giulianelli, Andreas Opedal, and Ryan Cotterell. Generalized measures of anticipation and responsivity in online language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 11648–11669. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.682. URL https://aclanthology.org/2024.findings-emnlp.682/.

Mario Giulianelli, Sarenne Wallbridge, Ryan Cotterell, and Raquel Fernández. Incremental alternative sampling as a lens into the temporal and representational resolution of linguistic prediction. *Journal of Memory and Language*, 148:104715, 2026. doi: 10.1016/j.jml.2025.104715. URL https://doi.org/10.1016/j.jml.2025.104715.

Adam Goodkind and Klinton Bicknell. 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*, pp. 10–18, 2018. doi: 10.18653/v1/W18-0102. URL https://aclanthology.org/W18-0102/.

Mark Gotham, Gianluca Micchi, Néstor Nápoles López, and Malcolm Sailor. When in Rome: A meta-corpus of functional harmony. *Transactions of the International Society for Music Information Retrieval*, 6(1):150–166, 2023. doi: 10.5334/tismir.165.

John Hale. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*, 2001. URL https://aclanthology.org/N01-1021/.

John Hale. Uncertainty about the rest of the sentence. *Cognitive Science*, 30(4):643–672, 2006. doi: 10.1207/s15516709cog0000_64.

John Hale, Chris Dyer, Adhiguna Kuncoro, and Jonathan Brennan. Finding syntax in human encephalography with beam search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2727–2736. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1254. URL https://aclanthology.org/P18-1254/.

Daniel Harasim, Martin Rohrmeier, and Timothy J. O'Donnell. A generalized parsing framework for generative models of harmonic syntax. In *Proceedings of the 19th International Society for Music Information Retrieval Conference*, pp. 152–159, 2018. doi: 10.5281/zenodo.1492367.

Steffen A. Herff, Daniel Harasim, Gabriele Cecchetti, Christoph Finkensiep, and Martin Rohrmeier. Hierarchical syntactic structure predicts listeners' sequence completion in music. In *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*, pp. 903–909, 2021.

Steffen A. Herff, Leonardo Bonetti, Gabriele Cecchetti, Peter Vuust, Morten L. Kringelbach, and Martin A. Rohrmeier. Hierarchical syntax model of music predicts theta power during music listening. *Neuropsychologia*, 199:108905, 2024. doi: 10.1016/j.neuropsychologia.2024.108905.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4129–4138, 2019. doi: 10.18653/v1/N19-1419.

Kuan-Jung Huang, Suhas Arehalli, Mari Kugemoto, Christian Muxica, Grusha Prasad, Brian Dillon, and Tal Linzen. Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, 137:104510, 2024. doi: 10.1016/j.jml.2024.104510.

David Huron. *Sweet Anticipation: Music and the Psychology of Expectation*. MIT Press, 2006.

Frederick Jelinek and John D. Lafferty. Computation of the probability of initial substring generation by stochastic context-free grammars. *Computational Linguistics*, 17(3):315–353, 1991. URL https://aclanthology.org/J91-3004/.

Stefan Koelsch, Martin Rohrmeier, Renzo Torrecuso, and Sebastian Jentschke. Processing of hierarchical syntactic structure in music. *Proceedings of the National Academy of Sciences*, 110 (38):15443–15448, 2013. doi: 10.1073/pnas.1300272110.

Carol L. Krumhansl and Edward J. Kessler. Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. *Psychological Review*, 89(4):334–368, 1982. doi: 10.1037/0033-295X.89.4.334.

Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, and Kentaro Inui. Lower perplexity is not always human-like. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5203–5217. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.405. URL https://aclanthology.org/2021.acl-long.405/.

Kenichi Kurihara and Taisuke Sato. Variational Bayesian grammar induction for natural language. In *Grammatical Inference: Algorithms and Applications*, Lecture Notes in Computer Science, pp. 84–96, 2006. doi: 10.1007/11872436_8.

Fred Lerdahl. *Tonal Pitch Space*. Oxford University Press, 2001.

Fred Lerdahl and Ray Jackendoff. *A Generative Theory of Tonal Music*. MIT Press, 1983.

Fred Lerdahl and Carol L. Krumhansl. Modeling tonal tension. *Music Perception*, 24(4):329–366, 2007. doi: 10.1525/mp.2007.24.4.329.

Roger Levy. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177, 2008. doi: 10.1016/j.cognition.2007.05.006. URL https://doi.org/10.1016/j.cognition.2007.05.006.

Longhai Li, Shi Qiu, Bei Zhang, and Cindy X. Feng. Approximating cross-validatory predictive evaluation in Bayesian latent variable models with integrated IS and WAIC. *Statistics and Computing*, 26(4):881–897, 2016. doi: 10.1007/s11222-015-9577-2. URL https://doi.org/10.1007/s11222-015-9577-2.

Tal Linzen and T. Florian Jaeger. Uncertainty and expectation in sentence processing: Evidence from subcategorization distributions. *Cognitive Science*, 40(6):1382–1411, 2016. doi: 10.1111/cogs.12274.

João Loula, Benjamin LeBrun, Li Du, Ben Lipkin, Clemente Pasti, Gabriel Grand, Tianyu Liu, Yahya Emara, Marjorie Freedman, Jason Eisner, Ryan Cotterell, Vikash Mansinghka, Alexander K. Lew, Tim Vieira, and Timothy J. O'Donnell. Syntactic and semantic control of large language models via sequential monte carlo. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/a2d537e69a6c6638a3630eef835f07de-Abstract-Conference.html.

Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054, 2020. doi: 10.1073/pnas.1907367117.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993. URL https://aclanthology.org/J93-2004/.

James A. Michaelov, Megan D. Bardolph, Cyma K. Van Petten, Benjamin K. Bergen, and Seana Coulson. Strong prediction: Language model surprisal explains multiple N400 effects. *Neurobiology of Language*, 5(1):107–135, 2024. doi: 10.1162/nol_a_00105. URL https://escholarship.org/uc/item/6x21g808.

Shikhar Murty, Pratyusha Sharma, Jacob Andreas, and Christopher Manning. Grokking of hierarchical structure in vanilla Transformers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 439–448, 2023. doi: 10.18653/v1/2023.acl-short.38.

Matthew J. Nelson, Imen El Karoui, Kristof Giber, Xiaofang Yang, Laurent Cohen, Hilda Koopman, Sydney S. Cash, Lionel Naccache, John T. Hale, Christophe Pallier, and Stanislas Dehaene. Neurophysiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences*, 114(18):E3669–E3678, 2017. doi: 10.1073/pnas.1701590114. URL https://doi.org/10.1073/pnas.1701590114.

Byung-Doh Oh and William Schuler. 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, 2023. doi: 10.1162/tacl_a_00548. URL https://aclanthology.org/2023.tacl-1.20/.

Art B. Owen. Monte carlo theory, methods and examples. Online book, 2013. URL https://artowen.su.domains/mc/.

Koyena Pal, Jiuding Sun, Andrew Yuan, Byron Wallace, and David Bau. Future lens: Anticipating subsequent tokens from a single hidden state. In *Proceedings of the 27th Conference on Computational Natural Language Learning*, pp. 548–560, 2023. doi: 10.18653/v1/2023.conll-1.37.

Marcus T. Pearce and Geraint A. Wiggins. Expectation in melody: The influence of context and learning. *Music Perception*, 23(5):377–405, 2006. doi: 10.1525/mp.2006.23.5.377.

Marcus T. Pearce and Geraint A. Wiggins. Auditory expectation: The information dynamics of music perception and cognition. *Topics in Cognitive Science*, 4(4):625–652, 2012. doi: 10.1111/j.1756-8765.2012.01214.x.

Tiago Pimentel and Clara Meister. How to compute the probability of a word. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18358–18375, 2024. doi: 10.18653/v1/2024.emnlp-main.1020. URL https://aclanthology.org/2024.emnlp-main.1020/.

Tiago Pimentel, Clara Meister, Ethan G. Wilcox, Roger P. Levy, and Ryan Cotterell. On the effect of anticipation on reading times. *Transactions of the Association for Computational Linguistics*, 11:1624–1642, 2023. doi: 10.1162/tacl_a_00603. URL https://doi.org/10.1162/tacl_a_00603.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Brian Roark. Probabilistic top-down parsing and language modeling. *Computational Linguistics*, 27 (2):249–276, 2001. doi: 10.1162/089120101750300526.

Brian Roark, Asaf Bachrach, Carlos Cardenas, and Christophe Pallier. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 324–333. Association for Computational Linguistics, 2009. URL https://aclanthology.org/D09-1034/.

Martin Rohrmeier. Towards a generative syntax of tonal harmony. *Journal of Mathematics and Music*, 5(1):35–53, 2011. doi: 10.1080/17459737.2011.573676.

Martin Rohrmeier. Musical expectancy: Bridging music theory, cognitive and computational approaches. *Zeitschrift der Gesellschaft für Musiktheorie*, 10:343–371, 2013. doi: 10.31751/724.

Adrian Staub and Charles Clifton, Jr. Syntactic prediction in language comprehension: Evidence from *either…or*. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32 (2):425–436, 2006. doi: 10.1037/0278-7393.32.2.425. URL https://doi.org/10.1037/0278-7393.32.2.425.

Andreas Stolcke. An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *Computational Linguistics*, 21(2):165–201, 1995. URL https://aclanthology.org/J95-2002/.

Xiangbin Teng, Pauline Larrouy-Maestri, and David Poeppel. Segmenting and predicting musical phrase structure exploits neural gain modulation and phase precession. *Journal of Neuroscience*, 44(30):e1331232024, 2024. doi: 10.1523/JNEUROSCI.1331-23.2024.

Barbara Tillmann, Jamshed J. Bharucha, and Emmanuel Bigand. Implicit learning of tonality: A self-organizing approach. *Psychological Review*, 107(4):885–913, 2000. doi: 10.1037/0033-295X.107.4.885.

Marten van Schijndel and Tal Linzen. Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science*, 45(6):e12988, 2021. doi: 10.1111/cogs.12988.

Ethan G. Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470, 2023. doi: 10.1162/tacl_a_00612. URL https://aclanthology.org/2023.tacl-1.82/.

Kevin Yang and Dan Klein. FUDGE: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3511–3535, 2021. doi: 10.18653/v1/2021.naacl-main.276.

Shunyu Yao, Binghui Peng, Christos Papadimitriou, and Karthik Narasimhan. Self-attention networks can process bounded hierarchical languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3770–3785, 2021. doi: 10.18653/v1/2021.acl-long.292.

Ryo Yoshida, Hiroshi Noji, and Yohei Oseki. Modeling human sentence processing with left-corner recurrent neural network grammars. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 2964–2973. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.235.

Stephen Zhao, Rob Brekelmans, Alireza Makhzani, and Roger Baker Grosse. Probabilistic inference in language models via twisted sequential Monte Carlo. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 60704–60748. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhao24c.html.

A PROBABILITY CONSTRUCTION AND ESTIMATOR CHECKS

Word-event probability. Each language prefix ends at a word boundary, so the next word must begin with whitespace. With q_0 the GPT-2 sampling distribution (end-of-sequence suppressed) and B the event that the continuation begins with whitespace,

$$q_E(c \mid h) = \frac{\mathbb{E}_{q_0}[\mathbf{1}_B \mathbf{1}\{\pi(\omega) = c\}]}{\Pr_{q_0}(B \mid h)}, \tag{4}$$

and the surface distribution is correspondingly $q = q_0(\cdot \mid B, h)$. This sums over all BPE spellings of a word rather than reading one softmax entry (Pimentel & Meister, 2024). Rare rated words are seldom sampled, so an additional sample of futures that begin with the rated word is drawn and combined with the ordinary sample by stratified estimation.

Built-in null. If every $G_\ell = 1$, then $\Psi_h \equiv 1$ for any H and d_h, because a normalized autoregressive sampling law returns its first-token marginal:

$$\Pr_q(\omega_1 = c \mid h) = q(c \mid h) \sum_{\omega_{2:H}} \prod_{j=2}^{H} q(\omega_j \mid h, \omega_{<j}) = q(c \mid h).$$

With finitely many futures this holds only up to sampling error, so in music, where q_E is the exact softmax, the two can differ for rare events. Independent first-token counts have binomial variance, and $\mathbb{E}\|\widehat{q}_N - q\|_2^2 = (1 - \sum_c q(c)^2)/N$. The reference must match the actual sampler: our GPT-2 cache suppresses EOS. For decoded words the identity instead concerns the pushforward under π: an initial token spelling `_mark` can extend to *marker*, so its softmax probability is not the probability of the closed word *mark* (Pimentel & Meister, 2024). Both arms use the same target-blind whitespace-boundary event B and word projection.

From scores to next-event probabilities. Equivalently to Eq. 3, $p_{\mathrm{syn}} = q_E \Psi_h$ mixes over lengths the next-event marginals of the grammar-tilted distributions $r_\ell(\omega \mid h) \propto q(\omega \mid h) G_\ell(\omega, h)$:

$$p_{\mathrm{syn}}(c \mid h) = q_E(c \mid h)\Psi_h(c) = \sum_{\ell=1}^{H} d_h(\ell)\, \frac{\mathbb{E}_q[\mathbf{1}\{\pi(\omega) = c\}\, G_\ell(\omega, h)]}{\mathbb{E}_q[G_\ell(\omega, h)]}. \tag{5}$$

Lengths are weighted by $d_h(\ell)$ rather than by $d_h(\ell) Z_\ell(h)$, as joint normalization over length and future would do. This modeling choice keeps the expected completion length separate from the scale of the music scores, which shrink as more rule probabilities are multiplied.

SNIS estimator. With N independent futures $\omega^{(i)} \sim q$, each length term of Eq. 5 is estimated as

$$\widehat{p}_{\mathrm{syn},\ell}(c \mid h) = \frac{\sum_{i=1}^{N} G_\ell(\omega^{(i)}, h)\, \mathbf{1}\{\pi(\omega^{(i)}) = c\}}{\sum_{i=1}^{N} G_\ell(\omega^{(i)}, h)}. \tag{6}$$

Because the proposal is q itself, the importance weights are the scores; weighting again by q would count frequent futures twice. In language, futures that do not begin at a word boundary receive zero weight. The estimator has finite-sample ratio bias but is consistent under the usual support and moment conditions (Owen, 2013). Every length had positive estimated support in the reported analyses.

Monte Carlo precision. In 100 paired resamplings of the language caches, with the complete behavioral procedure refitted each time, the Ours − Baseline contrast has SD .038 ELPD at $H = 10$ and .058 at $H = 2$, against sentence SEs of 1.392 and 2.032.

B GRAMMAR AND SURFACE IMPLEMENTATIONS

Table 3 summarizes how the construction of Section 3 is instantiated in music and language.

Table 3: One probability operator, two structural instantiations. Neither grammar is fitted to human ratings.

Component	Music	Language
Surface event	One chord symbol	Complete word/punctuation event
Grammar terminal	Chord root and quality	Part-of-speech tag
Open goal	Scale degree in a local key	Phrase category (e.g., VP)
Score G_ℓ	Probability of completing the goals	Progress toward completing the goals
Gate	Certainty about the first goal	How well the first goal predicts the next tag

B.1 CHORD GRAMMAR: RULES AND ESTIMATION

A category $X_{d,\kappa}$ denotes degree d in local key $\kappa = (r, m)$. Degree labels are indices, not chord qualities. Roots use

$$\mathrm{root}(d, \kappa) = (r + \delta_m(d)) \bmod 12, \quad \delta_{\mathrm{maj}} = (0, 2, 4, 5, 7, 9, 11), \quad \delta_{\mathrm{min}} = (0, 2, 3, 5, 7, 8, 10). \tag{7}$$

Let $\mu(d, \kappa) = (\mathrm{root}(d, \kappa), \eta_m(d))$. The fixed target modes are

$$\eta_{\mathrm{maj}} = (\mathrm{maj}, \mathrm{min}, \mathrm{min}, \mathrm{maj}, \mathrm{maj}, \mathrm{min}, \mathrm{min}),$$
$$\eta_{\mathrm{min}} = (\mathrm{min}, \mathrm{min}, \mathrm{maj}, \mathrm{min}, \mathrm{maj}, \mathrm{maj}, \mathrm{maj}).$$

The minor fallbacks for diminished degrees are modeling conventions, not established modulations; they are broader than the restriction in Rohrmeier (2011). For $d \neq \mathrm{I}$, tonicization and dominant preparation are

$$X_{d,\kappa} \to X_{\mathrm{I},\mu(d,\kappa)}, \qquad X_{d,\kappa} \to X_{\mathrm{V},\mu(d,\kappa)} X_{d,\kappa}. \tag{8}$$

Remaining schemas, adapted from Harasim et al. (2018), are

$$\begin{aligned}
X_{d,\kappa} &\to X_{d,\kappa} X_{d,\kappa}, & X_{d,\kappa} &\to X_{d+4,\kappa} X_{d,\kappa} \quad (d \neq \mathrm{IV}),\\
X_{\mathrm{I},\kappa} &\to X_{\mathrm{IV},\kappa} X_{\mathrm{I},\kappa}, & X_{\mathrm{I},(r,m)} &\to X_{\mathrm{I},(r,\bar{m})},\\
X_{\mathrm{I},(r,\mathrm{maj})} &\to X_{\mathrm{VI},(r,\mathrm{maj})}, & X_{\mathrm{I},(r,\mathrm{min})} &\to X_{\mathrm{III},(r,\mathrm{min})},\\
X_{\mathrm{II},\kappa} &\to X_{\mathrm{IV},\kappa}, & X_{\mathrm{V},\kappa} &\to X_{\mathrm{VII},\kappa},\\
X_{\mathrm{V},(r,m)} &\to X_{\mathrm{V},(r+s \bmod 12,m)}, & s &\in \{3, 6, 9\}.
\end{aligned}$$

Degree addition is modulo seven with I indexed by zero; $\bar{m}$ is the opposite mode. Each degree/mode distribution jointly covers structural rules and ten unrestricted chord-form emissions: major, major seventh, major sixth, dominant seventh, minor, minor seventh, half-diminished seventh, diminished seventh, augmented and suspended. Probabilities are tied across tonic roots. Unlike theory-constrained termination, no degree-specific form exclusions apply. Natural-minor VII emits on the subtonic; uppercase V does not force a major chord. The grammar omits inversions, voice leading, meter and cadence strength.

Training expresses each fragment relative to the annotated tonic of its region. The root category of a training fragment is drawn from a fixed uniform prior over the fourteen degree–mode categories of that tonic (seven degrees in major and in minor), so fragments need not begin on the tonic; this prior is not learned. Prediction uses a fixed equal start mixture over $X_{\mathrm{I},(0,\mathrm{maj})}$ and $X_{\mathrm{I},(0,\mathrm{min})}$; the tonic itself is not inferred over all keys. Evaluation inherits source-aligned tonal coordinates, constant within each analyzed excerpt. Latent local keys inside parses are distinct from this supplied reference and do not establish a perceptual modulation. Adjacent duplicates collapse in training, and fragments are not truncated. Evaluation preserves response positions, including 34 adjacent pairs merged by grammar projection.

Trees are latent. A single grammar is estimated from the 709 corpus songs that are not sources of any Cheung stimulus: 6,983 fragments with 68,919 chords after collapsing adjacent duplicates. We use a fully factorized Dirichlet variational posterior Q over the tied rule distributions (Kurihara & Sato, 2006). Each fragment, one occurrence of an annotated song section within one tonic region, is a complete, finite observation: its latent derivation, rooted in the fixed fragment prior, must generate exactly the observed chords. One deterministic full-batch coordinate update first recomputes the

optimal local tree distribution for every training fragment from the current expected log rule weights, using exact inside expected counts, and then sets

$$Q(z_i) \;\propto\; \exp\{\mathbb{E}_{Q(\theta)}[\log p_\theta(x_i, z_i)]\}, \qquad \nu_r \leftarrow \alpha + \sum_{i \in \mathcal{D}} \mathbb{E}_{Q(z_i)}[n_r(z_i)]. \tag{9}$$

Here $\alpha = 1$ and $\mathcal{D}$ contains every training fragment. Length-dependent microbatches only bound memory: their expected counts are accumulated at one fixed global state before the unit-step update, so they do not make the algorithm stochastic. Updates stop once the gain in the evidence lower bound falls below 10^{-5} nats per training chord in two consecutive updates, which happened after 128 updates; no behavioral or neural outcome enters training or stopping. Prediction uses normalized posterior-mean rule probabilities. The implementation checks support before training and count conservation and monotone bound improvement during training. After training it checks that the grammar is subcritical, with offspring spectral radius 0.960, so that derivations terminate with probability one, and that the 350 held-out fragments of the 30 source songs have finite scores. No gold trees, test songs or human behavioral or neural responses enter estimation.

B.2 Prefix charts and language adapter

Both domains use unit closure $R_U = (I - P_U)^{-1}$ and left-corner closure $R_L = (I - P_L)^{-1}$ on original categories. Music uses $P_L = R_U P_{\text{left}}$; language includes the first child of every production. Their spectral radii are below one for the reported grammars. Language's unit and left-corner radii are .015 and .30, and its termination is certified by extinction iteration. No category-changing grammar transform is applied. A span chart combines boundary-completion and split terms, closed under R_L. Normalizing by continuing-analysis mass treats the prefix as incomplete. A second chart and top-down adjoint yield one-goal and ordered two-goal masses, summing out deeper goals. The first open goal F has marginal $p(F = x \mid h) = p_1(x \mid h) + \sum_y p_2(x, y \mid h)$; this marginal and the total mass are checked.

Language uses the relative-frequency PCFG of 3,914 NLTK PTB trees: 72 nonterminals, 114 unary, 511 binary and 3,130 longer productions, without binarization, and unknown-word pseudo-count .05. Prefixes use lexical emissions. Future projection preserves structural rules but sums out lexical emissions; a word-form adapter provides independent soft POS evidence $p(t \mid w)$ with rare-word backoff. This posterior evidence reintroduces a tag-frequency prior and is an approximation, not a lexical likelihood or unsupervised discovery of syntax. The chart divides compatible mass into the four progress states of Section 3.5, using $G_\ell = \sum_s u_s m_s / \sum_s m_s$, $u = (1, 2, 2, 3)$, or zero when compatible mass vanishes. Averaged over events and over lengths weighted by d_h, about three quarters of the compatible mass still has a tracked goal open, which is why language scores progress rather than completion. The score measures progress rather than probability, so a rare compatible path scores the same as a common one; a future also scores zero once it runs past both completed goals. Music instead uses exact-yield inside probabilities (Eq. 2). Replacing yield probabilities by total length masses gives the completion-length distribution

$$d_h(\ell) \;\propto\; \sum_{g_1} p_1(g_1 \mid h)\, \Pr_{\mathcal{G}}\big(|\text{yield}(g_1)| = \ell\big) + \sum_{g_1, g_2} p_2(g_1, g_2 \mid h)\, \Pr_{\mathcal{G}}\big(|\text{yield}(g_1 g_2)| = \ell\big), \tag{10}$$

the probability under the grammar $\mathcal{G}$ that completing the tracked goals takes ℓ terminals, truncated at H and renormalized over $1 \le \ell \le H$. Before this renormalization, the retained mass at $H = 10$ averages 74.5% in music and 74.4% in language. Gate statistics use the first-goal posterior; language additionally uses $P(T \mid F = x) = [R_L E]_{x,t}$, with E the preterminal-emission matrix. The language gate's threshold of four prefix terminals was chosen during development by the log loss of the actual next words on training sentences, never by ratings.

Worked open goals. Take $S \to \text{SUBJ VP STOP}$, $\text{SUBJ} \to \text{Alice}$, $\text{ADV} \to \text{quickly}$, $\text{STOP} \to .$, and equiprobable $\text{VP} \to \text{runs} \mid \text{ADV VP}$. After *Alice*, goals are (VP, STOP). *runs* . completes VP and then STOP (utilities $2, 3$); *quickly runs* leaves VP open and then completes it $(1, 2)$. With equal proposal mass and $H = 2$, the truncated duration mass is all at two, so structural weighting changes $\Pr(\text{runs})$ from $1/2$ to $3/5$. Exact completion instead scores these paths $1/2$ and zero. This toy is not a PTB parse: its one-terminal prefix has zero maturity gate, so p_{hyb} equals q_E.

B.3 SURFACE MODELS AND EXECUTION

Music uses a four-layer, four-head pre-normalization Transformer (embedding 256, feed-forward 1,024) and a three-layer LSTM (embedding 256, hidden 384), context up to 128 tokens, and next-chord cross-entropy. Ten duplicate-grouped folds partition 739 Billboard songs; the grammar is instead estimated once from the 709 songs that are not sources of any stimulus (Appendix B.1). Training duration is chosen within the training partition before reinitialization and full-partition fitting. Each test excerpt uses the fold excluding its source song. IDyOM combines long-/short-term models by entropy-weighted geometric pooling; hypothetical events never update counts. q_E and p_{hyb} receive the same uniform smoothing mass, $36/1060$ spread over the 739 symbols, which keeps surprisal finite and was not tuned on human data. Language uses frozen pretrained GPT-2 and cached proposals.

Both domains track two goals. Music uses $H = 10$ chords, $N = 16{,}384$ and three streams (seeds 11, 22, 33); language uses $H = 2$ PTB terminals, $N = 4{,}096$ ordinary and $1{,}024$ target-conditioned futures, with a sweep over $H = 1, \ldots, 10$ as sensitivity. PTB terminals are not words or BPE tokens: 240 target words map to multiple terminals. Language starts with 80 generated BPE tokens (after the forced prefix for target draws), allowing up to 4,096 additional tokens. Censored paths retain reached-length scores and zero at unreached lengths, without changing denominators. One ordinary path remains censored; every positive-duration event/length cell has recorded support. Per-prefix grammar scoring takes about .5 seconds for music's three streams and 2.6 seconds for language on one GPU.

C RESPONSE MODELS AND PREDICTIVE SCORING

C.1 MUSIC: REPRESENTED-SONG AND UNSEEN-SONG TARGETS

Surprise data from 25 participants yield 2,018 musician-group means at 1,009 non-initial positions across 30 songs (Cheung et al., 2024). We follow the response model in the authors' released analysis code, replacing only its focal IC predictor. The Gaussian identity-link population design is $\mathrm{VAL} + \mathrm{ARO} + \mathrm{musician} * \mathrm{IC}$, with VAL and ARO the valence and arousal ratings; all six coefficients including intercept vary jointly by song with covariance $\Sigma = \mathrm{diag}(\tau)\Omega\,\mathrm{diag}(\tau)$. These song effects are partially pooled rather than estimated independently. Continuous variables use full-panel mean/SD standardization; musicianship is binary. Within 60 song-by-musicianship series, residuals satisfy $r_i - \rho r_{\mathrm{prev}(i)} \sim \mathcal{N}(0, \sigma^2)$, with $r_i \sim \mathcal{N}(0, \sigma^2)$ at series start (nonstationary initialization). Thus ρ accounts for serial dependence between adjacent chord ratings, while shared song effects link the two musicianship groups.

Pleasure from 39 participants yields 1,009 position means on the same stimuli. Its fixed design is $\mathrm{IC} * \mathrm{ENTROPY} + \mathrm{valence} + \mathrm{arousal} + \mathrm{dissonance} + \mathrm{spectralcentroid} + \mathrm{spectralcomplexity}$, all nine coefficients varying by song, with 30 AR series. Both IC and entropy change between arms; stimulus reuse is not independent replication. Both outcomes use $\mathcal{N}(0, 1)$ fixed-effect priors, half-Cauchy$(0, 1)$ scale priors, LKJ(2) correlations and uniform $\rho \in (-1, 1)$. Four chains retain 2,500 of 5,000 iterations each, with adaptation target .99 and maximum tree depth 10. Sampler and predictive-scoring diagnostics of every reported fit are summarized in Appendix E.

The released analysis applies `brms` LOO to the same fitted model. We retain that response model but, because its observations are correlated, make our predictive target explicit and evaluate the conditional density of a missing response given the remaining responses in its song. For this conditional response-LOO, integrate song effects. Let A_g have diagonal one and subdiagonal $-\rho$ within each series; $\mu_g = X_g\beta$ and

$$V_g = Z_g \Sigma Z_g^\top + \sigma^2 A_g^{-1} A_g^{-\top}, \qquad Q_g = V_g^{-1},$$
$$f_i(\psi) = \mathcal{N}\left(y_i;\; y_i - \frac{[Q_g(y_g - \mu_g)]_i}{[Q_g]_{ii}},\; [Q_g]_{ii}^{-1}\right).$$

The last argument is variance; the mean cancels y_i. PSIS uses inverse-density ratios $1/f_i(\psi)$ from full-posterior global draws $\psi = (\beta, \Sigma, \sigma, \rho)$ (Bürkner et al., 2021; Li et al., 2016). This predicts an omitted response given the other responses in its song, not an unseen song. Contributions with Pareto $k > 0.7$ or pointwise MCSE > 0.03 receive exact held-out refits that integrate the missing response

while preserving AR indices; priors and preprocessing stay fixed. The Pareto-k, MCSE and refit diagnostics of the reported fits are summarized in Appendix E.

Paired event differences δ_i are averaged across streams first. Song-cluster SE is $\sqrt{30\,\mathrm{Var}_g(\sum_{i\in g}\delta_i)}$, using sample variance; bootstrap intervals resample songs. This is stimulus dispersion, not computation or selection uncertainty. The separate whole-song holdout fits 29 songs by Gaussian maximum likelihood and jointly predicts the thirtieth, with training-only scaling, the same fixed effects, and random intercept plus each arm's own IC slope. Every whole-song fit converged. Its ELPD is a different target.

C.2 Language and secondary outcomes

Language uses deterministic ten-fold sentence-held-out validation with training-only standardization and frequency imputation. X_s contains intercept, position, SUBTLEX log frequency, missing-frequency indicator, word length and focal IC; Z_s contains intercept and IC. The training covariance is $V_s = Z_s\Sigma Z_s^\top + \sigma^2\Gamma_s(\rho)$, with unrestricted 2×2 Σ and $[\Gamma_s]_{ij} = \rho^{|\mathrm{position}_i - \mathrm{position}_j|}$. Profiled maximum likelihood uses two starts and $\rho = .95\tanh\theta$. With $C_\beta = (\sum_{\text{train s}} X_s^\top V_s^{-1} X_s)^{-1}$, the held-out density is

$$y_s \sim \mathcal{N}\Big(X_s\widehat{\beta},\, Z_s\widehat{\Sigma}Z_s^\top + \widehat{\sigma}^2\Gamma_s(\widehat{\rho}) + X_s C_\beta X_s^\top\Big).$$

This integrates unseen-sentence effects and approximate fixed-effect uncertainty, conditional on fitted covariance and predictors. Natural-log joint densities are not independent word scores or exact sentence LOO. Same-fold response Jacobians cancel in contrasts. SEs and 10,000-draw percentile intervals use 205 sentences; cache-resampling precision is evaluated separately (Appendix A). Duration selection uses five sentence-grouped inner folds over 27 rules. Neither removes earlier dataset-level development.

The music EEG data (Teng et al., 2024) contain 62-channel recordings from 29 listeners who heard Bach chorales at three tempi. We use the original-order trials of nine chorales, excluding one whose audio does not match its reference score, which leaves 772 of 783 participant-trials after omitting, in every arm, trials with at most 500 artifact-free samples. Ocular artifacts are removed by ICA learned from non-music calibration periods. The 576 chord events come from expert Roman-numeral analyses (Gotham et al., 2023) aligned to the rendered beats at 100 Hz; labels follow the audio where it departs locally from the score, and a sensitivity variant keeps the score's labels. Two factorized LSTM surface models predict them: one trained in ten folds on Billboard and one adapted to 339 expert-annotated Bach chorales that exclude the stimuli and their tune families. Both use the chord grammar, $H = 10$, gate and smoothing of the main analysis, with probabilities averaged over three streams of 16,384 futures before IC and predictive entropy are computed from the root-and-quality marginal (primary) and from the root and full-chord marginals. The TRF regresses each channel on 0–600 ms lags of the cochlear envelope and of beat and chord-onset impulses (together, the acoustic features), plus IC and entropy impulses at chord onsets. Nine outer folds hold out all tempi of one chorale; four training-only validation chorales tune separate acoustic and chord penalties, including omission of the chord block. Held-out correlation r is averaged over channels and trials with chorales weighted equally, and participant-level paired differences receive 10,000-draw bootstrap intervals and Holm correction within each family of contrasts. Additional language outcomes and their limits appear below.

D Controls, sensitivity, and transfer

The analyses below retain frozen model predictions and report both supportive and limiting evidence. Apart from the LSTM controls, whose design was fixed before the results were computed, they are consistency, sensitivity and secondary analyses on the same materials. Numerical checks are summarized in Appendix E.

D.1 Music: alternatives to dynamic completion

Conditional response-LOO controls use all 1,009 events and three proposal streams (Table 4). The LSTM panel was fixed before the results were computed; One-step, Permuted and Entropy-matched

were repeated for the Transformer and IDyOM as consistency analyses whose design was recorded before they were run. Permuted preserves each prefix–length set of scores, One-step retains the same grammar and gate but bypasses sampled futures, and Constant gate replaces the event-specific gate by its outcome-blind mean, 0.620. The event-specific gate adds little over its mean (Ours − Constant gate, $+1.79 \pm 3.14$).

Across surface models, H2 holds throughout, while Ours − One-step ranges from about zero for the Transformer to a resolved gain for IDyOM. For IDyOM, Entropy-matched alone improves on Baseline ($+31.56 \pm 13.89$, interval $[6.74, 59.82]$), so part of its gain reflects the sharpness of its prediction; Ours still exceeds Entropy-matched by $+75.45 \pm 17.29$. For the Transformer, Permuted is slightly worse than Baseline (-4.59 ± 2.14, $[-9.02, -0.73]$).

Horizon. Varying H from 1 to 10 with the same grammar, gate, streams and response-LOO shows a step from $H = 1$ to $H = 2$ and a plateau thereafter. Ours − Baseline at $H = 1$ is $+2.54 \pm 31.05$, -1.21 ± 28.73 and $+44.16 \pm 34.29$ for the LSTM, Transformer and IDyOM, and at $H = 2$ it is $+62.36 \pm 15.98$, $+48.35 \pm 16.71$ and $+98.29 \pm 25.45$. No horizon exceeds $H = 10$ by more than $+2.93 \pm 4.12$ (Transformer, $H = 3$; interval $[-3.91, 11.86]$). One-step exceeds Ours with $H = 1$ by $+49.17 \pm 32.78$, $+48.36 \pm 26.15$ and $+36.25 \pm 26.30$. The next-chord distribution converges in the same way for all three surface models: its mean total variation from the $H = 10$ distribution is 0.335, 0.325 and 0.332 at $H = 1$, 0.067, 0.069 and 0.068 at $H = 2$, and at most 0.034 from $H = 3$ on, whereas Ours with $H = 10$ differs from Baseline by 0.194, 0.207 and 0.221.

Table 4: Controls under conditional response-LOO (ΔELPD $\pm$ song-cluster SE, with 95% paired song-bootstrap interval). Positive values favor the first condition. LSTM rows are the primary panel; Transformer and IDyOM rows are consistency analyses.

Condition	Condition − Baseline	Ours − condition
LSTM		
One-step	$+51.71 \pm 14.19$ $[24.46, 79.15]$	$+10.03 \pm 7.92$ $[-3.48, 26.88]$
Permuted	$+0.73 \pm 2.33$ $[-3.34, 5.69]$	$+61.02 \pm 15.20$ $[31.93, 90.30]$
Entropy-matched	$+3.23 \pm 6.16$ $[-8.00, 15.52]$	$+58.52 \pm 14.77$ $[29.60, 86.98]$
Constant gate	$+59.95 \pm 15.79$ $[29.30, 89.98]$	$+1.79 \pm 3.14$ $[-4.71, 7.51]$
Transformer		
One-step	$+47.15 \pm 12.69$ $[23.05, 72.05]$	$+0.03 \pm 5.15$ $[-10.06, 9.83]$
Permuted	-4.59 ± 2.14 $[-9.02, -0.73]$	$+51.78 \pm 13.83$ $[26.29, 79.34]$
Entropy-matched	$+4.38 \pm 9.88$ $[-15.12, 22.88]$	$+42.81 \pm 12.40$ $[19.03, 66.75]$
IDyOM		
One-step	$+80.41 \pm 18.77$ $[49.88, 121.63]$	$+26.61 \pm 9.76$ $[9.13, 46.42]$
Permuted	$+0.25 \pm 2.11$ $[-3.59, 4.57]$	$+106.76 \pm 25.62$ $[62.61, 160.64]$
Entropy-matched	$+31.56 \pm 13.89$ $[6.74, 59.82]$	$+75.45 \pm 17.29$ $[44.05, 110.54]$

For the static prior, b_f is the distribution of the first terminal generated from the grammar's predictive start distribution, an equal mixture of the major and minor home chord, computed in closed form through the unary and left-corner closures. Terminal mass is divided equally among the surface symbols that project to it, renormalized and smoothed like q_E. The single converged grammar supplies one prior for all stimuli. Both static alternatives retain the event's gate: Static mix is $(1 - a_h)q_E + a_h b_f$ and Static tilt is $(1 - a_h)q_E + a_h q_E b_f / \sum_c q_E(c) b_f(c)$. Thus the prior is static but the predictor remains contextual. A sensitivity analysis replaces b_f by the grammar's expected terminal frequency, $p_T \propto \boldsymbol{\lambda}^\top (I - A)^{-1} E$, with $\boldsymbol{\lambda}$ the start distribution, A the expected-offspring matrix over categories and E the category-to-terminal emission probabilities; it is finite because the spectral radius of A is 0.960. In this variant Static tilt reweights $q_E(c)$ by p_T of the terminal of c directly, as One-step does with the prefix-dependent law, so the tilt does not depend on how many surface symbols share a terminal; Static mix still divides p_T among them. Tonal covariates add tonic-root and dominant-root indicators and $-\log_2 b_f$ of the realized chord identically to the Baseline and Ours response models.

The auxiliary whole-song model uses an intercept and each arm's own IC slope, reducing covariance estimation with 29 training songs. Its scores do not replace primary response-LOO (Table 5). Ours improves over Baseline for every surface model. Static mix scores below Baseline and Static tilt

above Baseline under both priors, but each interval excludes zero only under p_T; Ours is distinguished from Static tilt only under p_T. Tonal covariates reduce Ours − Baseline from $+50.27$ to $+22.22$, with an interval that still excludes zero.

Table 5: Auxiliary whole-song holdout: 29 training songs predict the held-out song, scores averaged over the three sampling streams (ΔELPD $\pm$ song-cluster SE, 95% paired song-bootstrap interval). LSTM unless stated.

Contrast	ΔELPD $\pm$ SE	95% interval
Ours − Baseline, LSTM	$+50.27 \pm 12.09$	$[27.03, 74.00]$
Ours − Baseline, Transformer	$+38.46 \pm 11.37$	$[17.30, 61.00]$
Ours − Baseline, IDyOM	$+77.78 \pm 18.32$	$[45.06, 116.09]$
One-step − Baseline	$+38.46 \pm 10.36$	$[18.90, 58.68]$
Ours − One-step	$+11.81 \pm 6.48$	$[0.68, 25.46]$
One-step − Baseline, Transformer	$+34.11 \pm 10.20$	$[14.29, 54.17]$
Ours − One-step, Transformer	$+4.34 \pm 4.03$	$[-3.40, 12.11]$
One-step − Baseline, IDyOM	$+58.37 \pm 13.31$	$[35.25, 86.57]$
Ours − One-step, IDyOM	$+19.41 \pm 6.97$	$[6.90, 33.70]$
Primary static prior b_f (first terminal)		
Static mix − Baseline	-24.22 ± 15.52	$[-54.38, 5.32]$
Static tilt − Baseline	$+28.68 \pm 15.67$	$[-1.26, 59.64]$
Ours − Static mix	$+74.49 \pm 15.89$	$[46.52, 107.43]$
Ours − Static tilt	$+21.59 \pm 15.67$	$[-7.12, 53.37]$
Sensitivity prior p_T (expected terminal frequency)		
Static mix − Baseline	-26.80 ± 13.20	$[-52.20, -1.11]$
Static tilt − Baseline	$+30.21 \pm 11.22$	$[9.64, 52.84]$
Ours − Static mix	$+77.07 \pm 15.29$	$[49.18, 107.87]$
Ours − Static tilt	$+20.06 \pm 10.23$	$[0.90, 40.70]$
Ours − Baseline with tonal covariates	$+22.22 \pm 9.24$	$[4.77, 40.63]$

Worked and failure examples. Figure 3 traces one event through the LSTM computation: position 24 of song 22 (C major), where the progression has alternated between I and IV and listeners heard Dm7. The first open goal is I in C major (posterior 0.66), and the gate is $a_h = 0.74$. Of the 16,384 futures in one stream, 5,491 begin with C, which can complete this goal at once; they receive 2.16 times the uniform weight on average, against 0.91 for futures beginning with G, 0.13 for G7 and 0.010 for Dm7, which cannot complete the goal in one chord and has low completion probability at longer lengths under the learned grammar. Ours therefore raises the probability of C from 0.32 to 0.60 and lowers that of Dm7 from 0.016 to 0.004. The IC of the played Dm7 rises from 5.95 to 7.86 bits, from the 67th to the 74th percentile of all events, toward the 88th percentile of listeners' surprise.

The largest failure, selected after observing outcomes as the event with the smallest Ours − Baseline contribution (-2.35 nats over both group means), is position 33 of song 27 (A♭ major), where E♭7 has resolved to A♭ and listeners heard D♭ minor. The first open goal is I in E♭ major (posterior 0.39), and Ours raises the IC of D♭ minor from 9.66 to 10.71 bits (85th to 86th percentile), whereas listeners placed it at the 62nd percentile of surprise. Ours − Baseline is negative at 350 of the 1,009 events.

D.2 Language: horizon, gate, and comparator sensitivity

Horizon H truncates a completion-length mixture in PTB terminals, not a fixed word count. The ten-point sweep peaks at $H = 2$ (Figure 4); five sentence-grouped inner folds select it from 27 frozen duration rules in every outer fold. Ours ($H = 2$) exceeds Ours ($H = 1$) by 2.052 ± 0.999 ELPD, interval $[0.086, 3.955]$, but the $H = 1/H = 2$ IC correlation is .99988 and their joint increment is unresolved, with interval $[-4.913, 14.468]$. Nesting protects outer sentences during selection but does not remove dataset-level development.

Compared directly with the corresponding additive models (Table 6), Ours gains 4.768 ± 2.356 ELPD over Baseline plus released PSG tag surprisal, 8.569 ± 3.004 over Baseline plus PCFG next-terminal surprisal, 6.136 ± 2.657 over Baseline plus most-likely-tag probability, and 6.356 ± 2.367 over Baseline plus tag agreement. These comparisons establish the reported result only for the four tested scalar readouts.

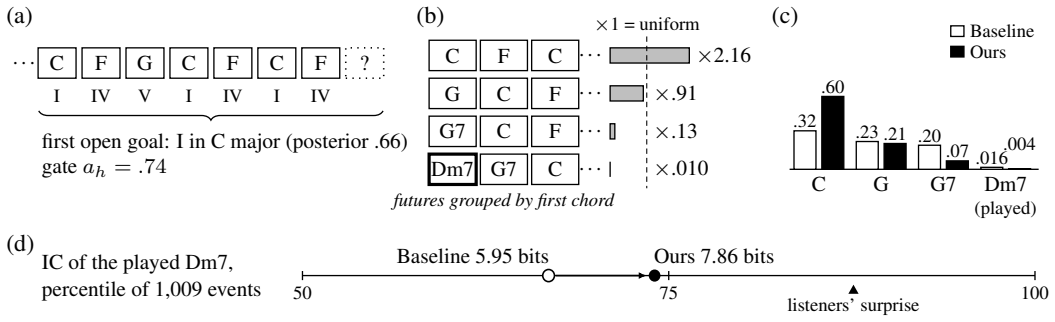


Figure 3: **Worked example** (LSTM, song 22, position 24). (a) The last seven chords of the prefix, the first open goal and the gate. (b) Sampled futures of one stream, grouped by first chord: mean weight relative to uniform, with each group's most frequent opening; the played chord is outlined. (c) Next-chord probability of the three chords most probable under Baseline and of the played chord, averaged over the three streams. (d) Percentile, among all 1,009 events, of the played chord's IC under Baseline and Ours and of listeners' mean surprise.

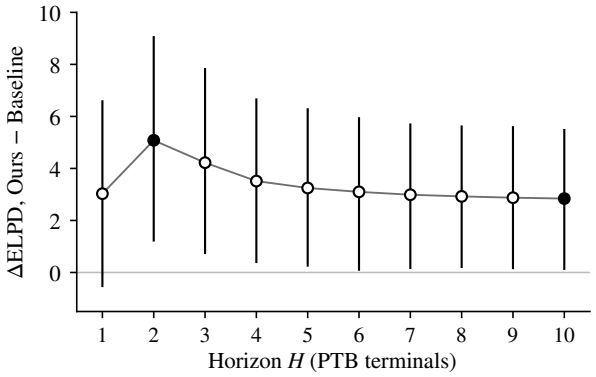


Figure 4: **Language horizon sensitivity.** Sentence-held-out rating prediction across horizons measured in PTB terminals. Points are Ours − Baseline with 95% paired sentence-bootstrap intervals; the selected $H = 2$ and the $H = 10$ reference are marked. This same-resource sensitivity is not an estimate of a human planning horizon.

Exact PCFG controls score the first PTB terminal; soft-POS compatibility is a potential rather than a normalized word probability. IAS comparisons use released Euclidean scores, not the unavailable article-final `cosine_std` values. The released IAS score adds 9.574 ± 4.509 ELPD to Baseline, and Ours adds 3.671 ± 4.910 after it. Under training-only nested IAS selection these values are -6.620 ± 6.262 and $+5.336 \pm 4.986$, respectively. Nested selection uses Gaussian OLS, whereas outer scoring uses sentence random slopes and AR residuals; its loss relative to fixed IAS concentrates in the three folds choosing another configuration. Both Ours-after-IAS increments remain unresolved.

D.3 Secondary outcomes

Pleasure ratings from 39 participants on the surprise stimuli (Cheung et al., 2024) were analyzed as a further transfer test, with IC, predictive entropy and their interaction in the released response model (Appendix C). Ours does not improve conditional response-LOO over Baseline for any surface model: $+5.20 \pm 6.96$ for the LSTM (interval $[-7.95, 18.85]$), -1.31 ± 9.24 for the Transformer ($[-19.74, 15.88]$) and -4.82 ± 8.15 for IDyOM ($[-20.88, 11.05]$). One LSTM response (song 28, chord 5) exceeded the 0.03 pointwise MCSE limit in all three streams; targeted exact refits replaced two of these scores, and the third refit missed the limit by 1.1×10^{-5}, so its sampling-based score is retained; using the refit instead changes the LSTM contrast by 0.007. In the music EEG, acoustic features alone give a mean held-out correlation of 0.0864, and adding chord IC and

Table 6: Language rating sensitivity. Intervals resample sentences; PSG, phrase-structure grammar. The upper row gives the $H = 10$ reference; the lower block asks whether adding a scalar syntax readout to Baseline is sufficient.

Contrast	ΔELPD $\pm$ SE	95% interval
Language ratings: Ours minus Baseline		
$H = 10$ reference	2.842 ± 1.392	$[0.099, 5.517]$
Language ratings: additive syntax predictors, (Baseline + predictor) minus Baseline		
Released PSG surprisal of tags	0.312 ± 1.927	$[-3.415, 4.161]$
PCFG next-terminal surprisal	-3.489 ± 2.106	$[-7.747, 0.715]$
PCFG most-likely-tag probability	-1.056 ± 1.532	$[-4.047, 1.937]$
PCFG tag agreement	-1.276 ± 1.045	$[-3.372, 0.747]$

entropy lowers it slightly. For the root-and-quality marginal, Baseline minus acoustic is -0.00039 (interval $[-0.00078, -0.00010]$) for the Billboard LSTM and -0.00055 ($[-0.00099, -0.00021]$) for the Bach-adapted LSTM, Ours minus acoustic is -0.00038 and -0.00048, and Ours minus Baseline is $+0.000005$ ($[-0.00011, 0.00014]$) and $+0.00007$ ($[-0.00013, 0.00035]$). Across both surface models, three marginals, two annotation variants, participant and piece units, and both r and predictive R^2, no chord arm exceeds acoustic features alone, and none of the 48 Ours $-$ Baseline contrasts survives Holm correction (smallest $p = 0.051$). For language N400, Ours minus Baseline is 1.942 ± 0.512 at $H = 2$, interval $[0.915, 2.907]$. At the $H = 10$ reference it is 0.687 ± 0.350, interval $[0.006, 1.363]$, and 0.673 ± 0.350, interval $[-0.007, 1.352]$, when the gate threshold is reselected within training sentences.

Ours ($H = 2$) minus Baseline on log self-paced/first-pass reading times is 0.094 ± 0.122/-0.239 ± 0.278; adding exact previous-word IC gives -0.256 ± 0.319/-0.321 ± 0.493. All four sentence-bootstrap intervals cross zero. Positive observed times use the rating folds and response family; spillover also controls previous-word lexical properties and requires an exact adjacent predecessor. Cloze counts are reconstructed from released responses and correction logs. The target-versus-rest score $k \log p + (n - k) \log(1 - p)$ gives Ours a positive but unresolved gain of $+128.61 \pm 71.66$ nats (sentence-bootstrap interval $[-9.38, 271.42]$, 135,923 responses); it does not evaluate the full cloze distribution. Sentence clustering does not capture dependence across items from shared respondents.

E Reproducibility and extended records

An artifact map, provided with the supplementary material, links the configurations, predictors, folds, seeds, scoring programs, checksums and complete comparison tables, and records the SHA-256 checksum of every artifact from which a reported number is taken.

The goal-certainty gate has mean 0.620 and range 0.342–0.858 over the 1,009 music events; it depends only on the prefix and is therefore identical for the three surface models. Before truncation at $H = 10$, the completion-length distribution retains on average 74.5% of its mass (median 74.1%), and every requested length has recorded support. The effective sample size of our model's importance weights, out of $N = 16{,}384$ futures, has median $4{,}117$, $3{,}679$ and $3{,}574$ for the LSTM, Transformer and IDyOM, with minima 18.0, 44.1 and 54.4. Recomputed Baseline IC matches each surface model's Baseline predictor within 10^{-6} bits, and recomputing Ours inside the control pipeline reproduces its probabilities within 5×10^{-8}. All 33 surprise response fits (Ours and its One-step, Permuted and Entropy-matched controls for the three surface models, and the LSTM Constant gate) have no divergent transitions or maximum-tree-depth hits, $\hat{R} \leq 1.005$ and minimum parameter ESS $1{,}669$; their integrated response-LOO scores have maximum Pareto $k = 0.361$ and maximum pointwise MCSE 0.005. The 81 additional horizon fits ($H = 1, \ldots, 9$) have no divergent transitions, $\hat{R} \leq 1.006$, maximum Pareto $k = 0.356$ and maximum pointwise MCSE 0.005. The nine pleasure fits have no divergent transitions or tree-depth hits, $\hat{R} \leq 1.005$ and minimum parameter ESS $1{,}179$; their integrated scores have maximum Pareto $k = 0.456$, and all pointwise MCSEs are at most 0.03 except the LSTM response described in Appendix D.3. Remaining limits include data access, environment portability, approximate computational uncertainty, and inference beyond observed language proposal support.