Measuring the Stories in Contemporary Songs

David Bamman, Sabrina Baur, Mackenzie Hạnh Cramer, Anna Ho, and Tom McEnaney

University of California, Berkeley

Abstract

Lyric poetry—the poetry of song—is often defined in opposition to narrative. In this work, we examine this relationship by carrying out an empirical study to measure the degree of *narrativity* present in contemporary songs, using a dataset of popular (Billboard Hot 100) and prestigious (Grammy-nominated) songs spanning 1960–2024. While we might expect the 1960s (with ballad-driven folk singers like Joan Baez, Bob Dylan and Simon & Garfunkel) to be a high-water mark for narrativity, we find the opposite: narrativity has been steadily increasing over this period, largely due to the rise of the strongly narrative genres of hip hop and rap. We also find that it is a marker of prestige for country music, with Grammy-award nominated "Best Country" songs displaying significantly higher narrativity rates than non-nominated songs from the same album.

Keywords: Narrativity, song lyrics, cultural analytics

1 Introduction

> And as I hung up the phone, it occurred to me
> He'd grown up just like me
> My boy was just like me
> (Harry Chapin, "Cat's in the Cradle")

Lyric poetry is the language of song; originally denoting accompaniment by lyre, but characterized by the first-person subject and attention to the personal experience of the poet, in contrast to the retelling of events in epic poetry.[1] While originally useful to differentiate Sappho from Homer, the rise of historical poetics upended a number of key claims in the tradition of lyric theory. Chief among those claims was Jonathan Culler's influential assertion, initially made in 1977 [8], and reiterated and expanded upon in subsequent articles and his 2015 book *Theory of the Lyric* [9], that lyric is a nonnarrative genre. In 1977 Culler suggested that one can "distinguish two forces in poetry, the narrative and the apostrophic"—a first-person subject addressing an absent interlocutor—and that "the lyric is characteristically the triumph of the apostrophic" [8, p. 66].[2] In

David Bamman, Sabrina Baur, Mackenzie Hạnh Cramer, Anna Ho, and Tom McEnaney. "Measuring the Stories in Contemporary Songs." In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 819–843. https://doi.org/10.63744/w9C0wDxmZTVt.

[1] "*Lyric* was from its inception a term used to describe a music that could no longer be heard, an idea of poetry characterized by a lost collective experience," writes Virginia Jackson in her definition of "lyric" for the *Princeton Encyclopedia of Poetry and Poetics* [15, p. 826]

[2] "Apostrophe resists narrative because its now is not a moment in a temporal sequence but a now of discourse, of writing" [8, p. 68]. While Culler's conflict with historical poetics seems to come down to the difference between transhistorical categories and historicist claims about the historical invention of labels like lyric, Culler himself admits that lyric became hegemonic in the 19th century as prose genres like the novel became the increasingly dominant homes for narrative writing [9, p. 76].

2015, Culler added, "The fundamental characteristic of lyric…[is] the iterative and iterable performance of an event in the lyric present, in the special 'now,' of lyric articulation …Fiction is about what happened next; lyric is about what happens now" [9, p. 226]. While Culler does not dismiss the existence of narrative poetry (he cites epic, didactic, and other poetic forms), and sometimes makes room for other historical uses of poetry, he firmly insists that "The notion of lyric as a genre, then, at bottom embodies a claim that *poetry* as a whole (which includes long narrative poems of various sorts) is in various ways a less useful category for thinking about poems than is *lyric*" [9, p. 89]. From the more limited notion of "apostrophic poetry" to the more totalizing understanding that lyric, rather than "*poetry* as a whole," best helps us theorize poems, Culler constructs a dividing line between the lyric genre and narrative.

We find an exemplary contrast to this position in Maureen McLane's *Balladeering, Minstrelsy, and the Making of Romantic Poetry* [30]. There McLane argues that ballads, or literary poems modeled on narrative songs with a stanzaic structure, were central to the invention of 19th century Romanticism, the historical moment and ideology that, perhaps more than any other, produced the poems and poetic theory most famous for restricting poetry to lyric as the voice crying out in the desert. This was the context that produced John Stuart Mill's often quoted phrase: "Poetry is overheard."[3] Coleridge and Wordsworth's 1798 *Lyrical Ballads* marks the transition point where lyric rises from the ashes of the ballad, burning off the tradition of collective storytelling for the modernity of singular expression (with all the Adornian caveats about how collective historical experience imprints itself upon even the most individualist lyric poem). That said, it was a slow burn, with the ballad's influence extending at least into the latter half of the 19th century, relevant still in the poetry of historical poetics' favorite poet: Emily Dickinson.[4]

The ballad has proven important to arguments that seek to historicize lyric not as fundamentally and structurally timeless (existing for all time and free of the unfolding time of narrative). On the other hand, Culler and his followers continue to assert that lyric is coherent, and its coherence depends, in part, on its rejection of narrative. This division has brought these theorists and others to ask after the influence of song in poetry, which leads us to wonder if there might be another lyric theory if we thought of the problem through *lyrics* theory.

For example, if the first-person subject of lyric emerged by both picking up and then throwing out the narrative baggage of the ballad, would the same hold true for music lyrics in the 20th century? Do song lyrics follow "lyric" and outright reject narrative? Does the tradition of singer-songwriters popularized in the 1960s with folk singers like Joan Baez, Bob Dylan, Simon & Garfunkel, Peter, Paul, and Mary, and others returning to the narrative form of the ballad establish a high-water mark for narrative in musical lyrics that has never returned?[5] In this article we seek to test whether "lyricization" and the totalizing genre of lyric has done away with narrative in music lyrics, or if lyrics might present us with a new reconciliation between lyric and narrative.

For decades, musicologists and other academic theorists have had surprisingly little to say about lyrics and narrative outside of scholars of country music[6] (to which we will return) and the oral tradition of folksong, which has charted the historical shift from ballad to lyric, including in the transmission of individual songs [1]. Just over a decade ago, Keith Negus wrote that "the popular song…has been almost entirely ignored in the vast literature on narrative" [38, p. 368]. A year later, Julia Simon, borrowing from Culler, argued that critics had implicitly aligned blues with lyric and ballad with epic poetry, but that the two "are drawn together in the efforts by the listener to

[3] See McLane [30], Stewart [53], and Mill [32] (later published as Mill [31]).

[4] Culler enlists Cristianne Miller's reading of Emily Dickinson [33] against Virginia Jackson's notion of "lyricization" [25], or the social invention of lyric, Jackson traces in her study of Dickinson: "In the early and mid-19th century United States, 'lyric' described any poetry that was not distinctly dramatic, epic, or narrative, that was harmonic or musical in its language, or that was conceived as song. Dickinson's poetry fits this model" (Miller quoted in Culler [9, p. 84].

[5] See Hampton [17].

[6] See Fox [13], Neal [37], and Stimeling [54], among others.

make lyric conform to understandable forms of narrative" [50, p. 52]. By 2019, Timothy Hampton, writing about Bob Dylan, would follow Simon's distinction (blues as lyric, ballad as narrative) to argue that Dylan maneuvers back-and-forth "within a kind of dialectic…to turn old forms to new purposes" [16, p. 161]. While these strategies within a song form might complicate what counts as narrative, what happens when we look across the history of popular song in the United States from the 1960s to the present, and attempt to trace the narrativity of lyrics?

2 Data

To do so, we first gather together a dataset of song lyrics, designed to capture both popularity and prestige: from Wikipedia we draw the Billboard Year-End Hot 100 singles from 1960-2024, which ranks singles based on (physical and digital) sales and airplay, including streaming. From `www.billboard.com` we also create a collection of popular songs by genre: Billboard year-end hot song charts for R&B/Hip Hop (2002–2024), Rock/Alternative (2009-2024), Country (2002-2024) and Rap (2013-2024). To build a collection focused on prestige, we identify all songs nominated for a Grammy *Song of the Year* over the period of 1960-2025. We draw lyrics for each song from `www.azlyrics.com`.

3 Framework

Our goal is to trace the narrativity of lyrics in contemporary song over a period of 65 years and explore its dynamics across a range of genres. Doing so analytically presents a challenge to measurement: how do we judge the degree of narrativity present in a song? This is fundamentally a question of operationalization [44; 45]—the act of transforming a theoretical construct into a measurable phenomenon.

A now growing body of research at the intersection of musicology and narrative has begun to explore the complex ways in which songs create narrative experiences for a listener by bringing together lyrics, music (including instrumentation, timbre, melody, and harmony) [41], the vocal staging [55], larger structural aspects such as sectional change [56], the record production [19], the persona adopted by the performer [36], and even the album art [41].

Nichols 2007 [41] provides one attempt at this, developing a 5-item Likert scale: songs at level 1 have "no story per se in the lyrics"; level-2 songs "contain elements of narrative discourse, but these are not reflected or supported in the (neutral) musical setting" and levels 3 and above offer increasing narrativity as the music and other modalities (including artwork) interact in complex ways. Songs like "Relax" (by Frankie Goes to Hollywood) and "I Want to Hold your Hand" (The Beatles) are examples of low-narrativity songs (level 1) in this system, while "Don't You Want Me" (Human League) show signs of narrativity (the song begins "You were working as a waitress in a cocktail bar …when I met you").

This operationalization is useful in providing a broad structure for the ways in which the holistic production of a song work together to inform the narrative experiences, but our goal is more narrow: to find the degree to which we see narrative embedded within the lyrics themselves, and how legible a story can be when only considering the words.

For this we require a finer-grained instrument narrowly focused on textual narratives. Modrow 2016 [35] provides one extensive theoretical system for the manual tagging of narrativity in songs that centers the concept of eventfulness and changes of state; in this work we draw as well on an increasing body of work using computational methods for narrative detection [4; 46; 47; 52]. We rely in particular on the work of Piper et al. 2021 [48] and Piper and Bagga [46; 47], who decompose narrativity across a range of genres into three axes (grounding in Herman 2009 [21]):

- **Agent**, the degree to which a passage "foregrounds the lived experience of particular agents."

- **Events**, the degree to which a passage is "organized around sequences of events that occur over time."
- **World**, the degree to which a passage "creates a world that I can see or feel."

This framework was developed in the context of a range of genres, including not only fiction and non-fiction, but poetry as well [48]. Hühn and Kiefer 2005 [23] converge on a related framework in their narratological analysis of lyric, emphasizing the role of events ("the temporal sequence of happenings" described in a poem) and mediation ("relating these happenings from a particular perspective") in defining narrative. We make similar adaptations as we interpret this framework for songs—for instance, counting changes of psychological state as events (as Hühn and Kiefer note, "in lyric poetry, happenings are frequently composed of mental or psychological processes") and attending to the agency of the first-person speaker.[7]

Like Piper and Bagga 2022 [47], we define each of these axes as a 5-point Likert scale, ranging from low (1) to high (5). In guiding our application of this framework to song, we pay particular attention to the *specificity* of description: while all songs could be interpreted to contain implied stories, we formalize this measure by rating the degree to which each of these elements are explicit in the lyrics; the greater the inference required to identify the actors involved, or the sequence of events in which they participate, the lower the judgment of narrativity. We codify this operationalization in a set of guidelines, which serve to circumscribe the boundaries of song narrativity and provide signposts to guide judgments of its presence in lyrics. The full annotation guidelines can be found in Appendix D.

This framework encodes the decision to treat individual songs in their entirety as the unit of analysis. This leaves open a range of alternative specifications that are equally interesting. For instance, to what degree does a narrative present within a song have internal structure [3] (as, for example, in Harry Chapin's *Cat's in the Cradle* [quoted in the epigraph], where the story describes a role reversal that unfolds over decades)? Likewise, what narrative techniques do entire albums use in their storytelling? Past work has explored the audio factors in track sequencing [39; 40], and many albums—ranging Pink Floyd's *The Wall* to Janelle Monáe's *The ArchAndroid*—have an explicit narrative arc that unfolds over the course of a sequence of songs. Our focus on the holistic narrativity of a song offers a natural starting point for analysis, and we leave to other work to explore these questions of narrative structure.

4 Annotation

Using the guidelines, three annotators, all co-authors, each independently provided judgments of narrativity for a total of 1,076 songs. From the dataset described in §2 above, we annotate the following:

- Grammy nominees for *Song of the Year*, 1960-2025, along with one other song from the same album.
- Billboard Year-End Hot 100, 1960–2024. We sample eight songs per year from these lists, sampling at random (but excluding any Grammy nominated songs of the year already identified above).

We only include songs for annotation with lyrics on `www.azlyrics.com`. Each annotator provided a rating along a 5-point Likert scale for each of the three axes defined above, yielding a

[7] "In lyric poetry, stories tend to differ from those of novels in that they are concerned primarily with internal phenomena such as perceptions, thoughts, ideas, feelings, memories, desires, attitudes, and products of the imagination that the speaker or protagonist ascribes to him- or herself as a story in a monological process of mental reflection, defining his or her individual identity by means of that story." [23, p. 18]

total of 9 annotations per song (3 for each axis by each of 3 annotators). In comparing the degree to which annotators agree with each other, we find an average deviation index of 0.56 (for reference, Piper and Bagga note an average deviation of 0.41 for reader judgments of narrativity in a range of fiction and non-fiction passages) and a chance-corrected measure, Krippendorff's α, of 0.46, speaking to the variability of narrative interpretations [34]. We find, however, strong pairwise agreement between annotators in the rankings of songs by their overall narrativity (average pairwise $\rho = 0.503$), suggesting individual differences in scaling. The composite narrativity score has higher agreement than any individual axis (agent $\rho = 0.348$, events $\alpha = 0.402$, world $\alpha = 0.481$); while the individual axes provide useful views on the narrative components (such as eventfulness) that comprise a story, annotators agree more on the fact of narrativity than on the specific component that makes it so.

We generate final scores for Agent, Events, and World as the average of all three independent annotations; any song with a standard deviation greater than 1 is then discussed to yield a final consensus rating. Figure 1 illustrates the distribution of these resulting annotations over the three categories; while many songs center the experiences of agents, fewer focus on sequences of events or build a world that is inhabitable by a listener.

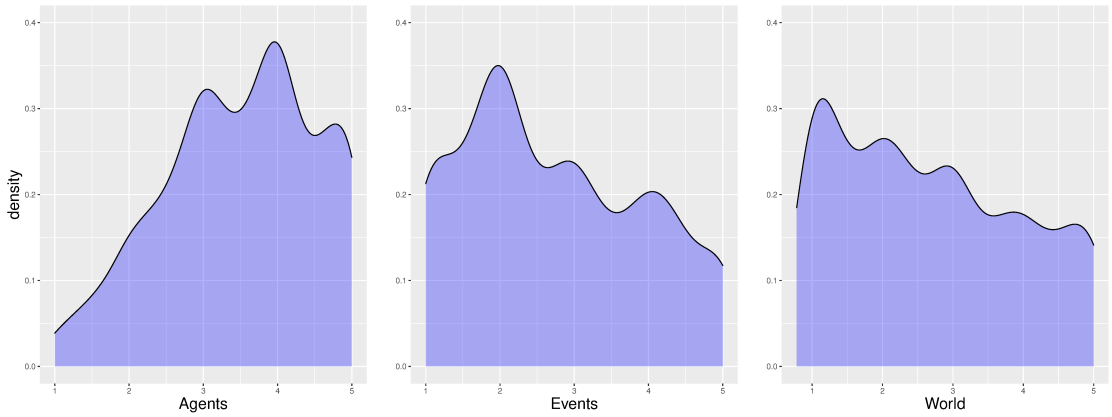

Figure 1: Distribution of annotations for agents, events, and world-building.

5 Models

The process described above created an annotated dataset of 1,076 songs. To explore the degree to which song narrativity can be modeled to enable larger-scale analysis, we use this data in a predictive machine learning framework, evaluating the capability of models to recreate these human judgments of narrativity [5]. We divided this dataset into partitions for training ($n = 646$), development ($n = 215$) and test ($n = 215$) and assessed the performance of a number of regression models for predicting a scalar value for each of the three dimensions using only information from the lyrics of the song. Each model is given access to the complete original lyrics of a song, preprocessing only to remove mentions of speaker turns (e.g. "[Taylor Swift:]") occassionally found when there are multiple singers in a song. For each model, we optimize hyperparameters (learning rate, ℓ_2 regularization strength) on development data using the `optuna` optimization library [2] over 50 optimization trials. We consider the following classes of models:

Minimal narrative features. Inspired by Piper and Bagga, we build a model using the five most informative features from that work that are able to predict narrativity in a number of (non-song) genres: the rate of past-tense verbs (VBD), nouns (NN) and present-tense verbs (VBZ),

concreteness (the average value of terms in the lexicon of Brysbaert et al. 2014 [7]), and the agenthood (the rate of animate entities participating as the subjects of verbs). Parts of speech are tagged and parsed for syntax using SpaCy in BookNLP. We train an ℓ_2-regularized linear regression on this dataset, using the development data to find the best ℓ_2 strength.

Featurized. We test several other feature combinations in ℓ_2-regularized linear regressions (using development data to again find the best ℓ_2 for each one). **1pp** includes only the rate of 1st-person pronouns (*I, me, my*); **POS + animacy + concreteness + 1pp** adds the animacy and concreteness features from the minimal narrativity feature set, along with rates for all parts of speech; **BoW + POS + animacy + concreteness + 1pp** additionally adds unigram bag-of-words features.

Masked LM. We explore four models pre-trained via masked language modeling objectives: BERT [11], RoBERTA [28], DeBERTa (V3) [20] and ModernBERT [57]. We train separate models for each of the three aspects of narrativity; to each base model we add a linear layer, minimizing the average mean squared error. We fine-tune all parameters on training data and use development data to find the best learning rate for each model.

Prompting models. We also assess the performance of three frontier LLMs: GPT 4.1 (OpenAI), Gemini 2.5 Pro (Google) and Claude Opus 4 (Anthropic). For all models, we prompt for 5-point Likert scale ratings on each of the three dimensions, providing three examples of input/output pairs; the full prompt can be found in appendix E.

5.1 Results

To assess performance, we generate a composite "narrativity" score as the average of Agent, Events, and World scores (for both the annotated data and all model predictions). We then measure the Spearman rank correlation coefficient (following Piper and Bagga [47]) between the true and predicted narrativity scores for all songs in the test data.

Table 1 illustrates the performance for all models, along with 95% bootstrap confidence intervals. We see that BERT-class models are the best performing (and practically indistinguishable from each other), with featurized bag-of-words models also competitive. LLMs and models without word features all struggle with this measurement. Performance by individual axis can be found in Table 5 (Appendix F).

Model	Spearman ρ
RoBERTa	0.840 [0.791-0.878]
DeBERTa-v3	0.837 [0.787-0.876]
BERT	0.828 [0.777-0.870]
ModernBERT	0.797 [0.735-0.846]
BOW + POS + animacy + concreteness + 1pp	0.776 [0.710-0.830]
`Gemini 2.5 Pro`	0.586 [0.485-0.674]
`GPT 4.1`	0.531 [0.410-0.633]
`Claude Opus 4`	0.530 [0.414-0.635]
POS + animacy + concreteness + 1pp	0.508 [0.396-0.603]
Minimal narrative features	0.427 [0.314-0.531]
1pp	-0.082 [-0.217-0.054]

Table 1: Model performance, along with 95% bootstrap confidence intervals.

We use perturbation-based methods [24] to investigate the most predictive terms for each narrative task in the best-performing RoBERTa model (for details, see Appendix G). The most frequent terms in local explanations for the agent task include not only a focus on the first-person (*I*) and *mirror* reflecting back on that first person, but also a focus on inebriation (*drunk, wine, wasted, drank*). The most common explanations for the event task include not only a focus on the past tense identified by Piper and Bagga (*said, planned, stole*), but also explicit temporal indicators (*tonight, yesterday, evening, tomorrow*). The world-building task includes explanations in specific places (*downtown, hall, town, mall, street, train, city*) and things *(coffee, wine)*.

6 Analysis

With a construct of narrativity in song defined, human judgments of narrativity for songs created, and predictive models able to reproduce those judgments with relatively high accuracy, we have the requisite pieces to turn back to our original question: how do we see narrativity expressed in the lyrics of contemporary songs?

We identify all songs from the Billboard Year-End Hot 100 from 1960–2024 and gather any available lyrics from `www.azlyrics.com` (a total of 5,745 songs). In this collection, the average song spans 423 tokens, with the most represented genres including pop (19.6%), contemporary R&B (16.5%), soul (8.8%), hip-hop (8.6%), pop rock (8.4%), country (8%), soft rock (7.9%), and dance pop (6.4%).

For each song, we use our best-performing model (RoBERTa) to predict its composite narrativity as the average of model predictions for Agent, Events, and World. Appendix B illustrates the most narrative songs within this collection; Appendix C provides a close reading of two of them (Taylor Swift's "All Too Well" and Ice Cube's "It was a Good Day").

6.1 Narrativity over time

How has the narrativity of songs changed over time? Figure 2 plots the average narrativity in the Billboard Year-End Hot 100 again over the period 1960-2024, along with 95% confidence intervals. We see a strong correlation over time ($\rho = 0.87, p < 0.001$), with an absolute increase in narrativity of 1.14 points from 1960 (2.58) to 2024 (3.72).[8]

This finding seems to challenge the assumptions we set out at the opening of this article, where, taking the familiar division between lyric (non-narrative) and ballad (narrative), we expected that the popular identification of the 1960s with the resurgence of traditional ballads by folk singers like Bob Dylan would have led to a sharp rise in the narrative drive of songs in the 60s. However, while Dylan appears on the Billboard charts with three different songs in the mid to late 60s, and covers of his songs by The Byrds and The Turtles chart as well, they're joined by only a small handful of other folk songs, and among them all, only Simon and Garfunkel's "Scarborough Fair" (#89 in 1968) is an arrangement of a traditional ballad. Instead of the ballad revivalists, Motown and the Beatles saturate the charts across the 60s.

The revolution in popular *narrative* songwriting and fandom, it seems, would take another thirty years, driven by another genre, associated not with the ballad, but with lyric: hip hop.

6.2 Genre

For songs in the Billboard Hot 100 (1960-2024), we examine the relationship between narrativity and genre by gathering fine-grained genre information for each song from Wikipedia. Table 2 presents the average narrativity among all genres attested by at least 200 songs; we see hip hop

[8] Additional robustness checks on model bias and annotator familiarity can be found in Appendix H.

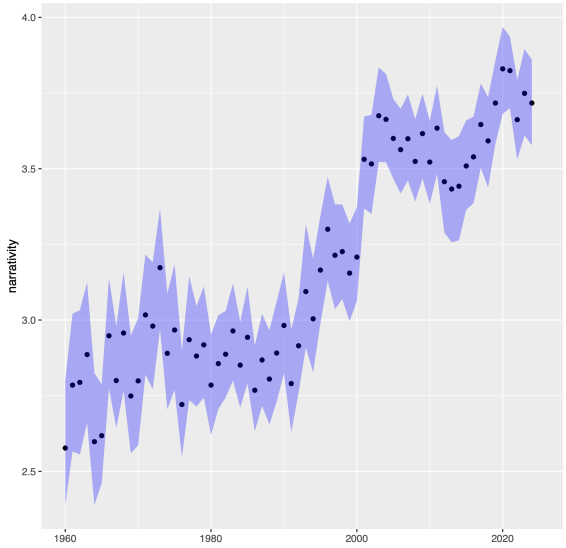

Figure 2: Narrativity over time, Billboard Year-End Hot 100 (1960-2024), with 95% confidence intervals.

clearly differentiated as the most narrative of genres, followed by country;[9] the least narrative of genres are disco and soul.

Genre	Narrativity	Agents	Events	World
Hip hop	4.01 [3.96-4.07]	4.47 [4.44-4.51]	3.84 [3.76-3.92]	3.72 [3.65-3.80]
Country	3.43 [3.35-3.51]	4.05 [4.00-4.10]	3.48 [3.38-3.59]	2.74 [2.62-2.87]
Contemporary R&B	3.19 [3.14-3.24]	4.06 [4.03-4.10]	3.18 [3.11-3.24]	2.32 [2.25-2.39]
Pop rock	3.12 [3.05-3.19]	3.93 [3.87-3.98]	3.12 [3.03-3.21]	2.32 [2.22-2.41]
Synth-pop	3.11 [3.00-3.22]	3.91 [3.82-3.98]	3.01 [2.86-3.15]	2.41 [2.26-2.56]
Rock	3.10 [3.00-3.20]	3.86 [3.80-3.93]	3.09 [2.96-3.23]	2.34 [2.21-2.48]
Dance-pop	2.97 [2.90-3.05]	3.81 [3.75-3.87]	2.83 [2.73-2.94]	2.27 [2.17-2.38]
Pop	2.95 [2.90-3.00]	3.81 [3.77-3.85]	2.92 [2.86-2.99]	2.11 [2.05-2.18]
Soft rock	2.91 [2.84-2.99]	3.79 [3.74-3.85]	2.92 [2.82-3.03]	2.03 [1.94-2.13]
Funk	2.86 [2.76-2.97]	3.65 [3.56-3.75]	2.64 [2.51-2.77]	2.29 [2.15-2.42]
Soul	2.72 [2.66-2.78]	3.66 [3.61-3.71]	2.64 [2.56-2.73]	1.86 [1.77-1.94]
Disco	2.71 [2.62-2.81]	3.54 [3.46-3.62]	2.55 [2.42-2.68]	2.05 [1.92-2.19]

Table 2: Narrativity by genre, along with 95% confidence intervals.

The dominance of hip hop and country provides important context for the rising trend in narrativity illustrated in figure 2. How much of this rise is due to the rise of hip hop itself? Figure 3 answers this by plotting the proportion of the Billboard Hot 100 songs that are comprised of both hip hop/rap (left) and country (right) songs. While country has seen a recent resurgence that may partially sustain narrativity today, hip hop increasingly dominated the charts during the narrative revolution in popular music starting in the mid-1990s—from fig. 2, we can see the 1990s brought a rapid ascent of narrativity that has persisted into the 21st century. The 1990s is where hip hop, not folk songs, mark the turning point for narrative in popular music.

[9] Modrow 2016 [35] finds a similar emphasis on narrativity and "super-narrativity" in hip hop and country within a collection of 78 German and English pop songs.

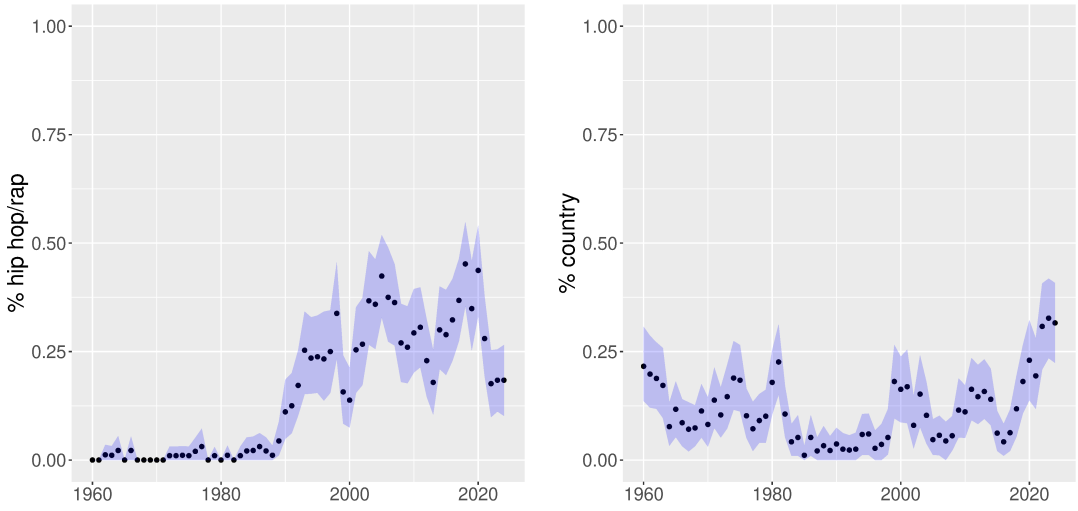

Figure 3: Proportion of songs on the Billboard Hot 100 that are hip hop/rap (left) and country (right). The rise in narrativity in fig. 2 is concurrent with the increasing charting of hip hop/rap songs on the Billboard Hot 100, and sustained by the renewed popularity of country in the 2020s.

That hip hop would drive narrative might seem surprising to critics, popular and academic, who argue that the genre, more than any other, marks the return of poetry to popular culture.[10] In their 2020 book *Rhymes in the Flow: How Rappers Flip the Beat*, authors Aurko Joshi and Macklin Smith confidently declare that "the lyric mode dominates rap" and "[n]arrative is relatively uncommon" [51, p. 149]. They in particular attribute this perceived rarity towards the genre's focus on the "spontaneous or spontaneous-seeming performance self," exemplified in first-person name drops, braggadocio, wordplay and aphorisms rounding out an artist's dramatic personae as well as foundational hip hop forms such as off-the-cuff freestyling, that preclude the "detachment" required for "sustained, song-length narrative[s]" and story arcs [51, p. 120]. More recently, Charlie Hankin takes up lyric theory from Culler and others to argue that hip hop freestyling is a "present-tense" exercise, a recursive work exemplified in what he names "raplove": "Without the time to elaborate a narrative arc, freestylers often rhyme about 'what is happening, usually in present tense.' Freestyling, in other words, is a poetics of epistrophic return, a turning about on 'lo que hay' (what there is)" [18, p. 50]. The implication that the fast-talking, improvisational subsets of hip hop are restricted not only by temporality but also time in their ephemeral performances overlooks the genre's world-building, chronicling, and characterizational narrativity. While these authors acknowledge minor narrative moments in hip hop, their work cannot conceive of hip hop's narrative influence.

While the charts above examine narrativity in the Billboard Top 100 (and see an explanation for that rise in the composition of the Top 100 itself), we can examine this same phenomenon in more detail within each genre by considering the Billboard year-end "Hot Song" charts for R&B/Hip Hop (2002–2024), Rock/Alternative (2009-2024), Country (2002-2024) and Rap (2013-2024)—lists again driven by popularity in sales and airplay but specific to each genre. Figure 4 illustrates this fuller narrative picture in genre over time. We can see that country, R&B/hip hop and rap have all been relatively stable in their individual narrativity over time (with rap showing the highest sustained levels), lending credence to the observation that changing narrativity within the Billboard charts is primarily due to the changing genre composition of the chart itself; but we also see strong

[10] See Hankin [18], Culler [10], Bradley [6] and Perry [43].

increases in narrativity *within* the genre of rock and (to a lesser extent) pop as well, reflecting a potential influence that cuts across genres.

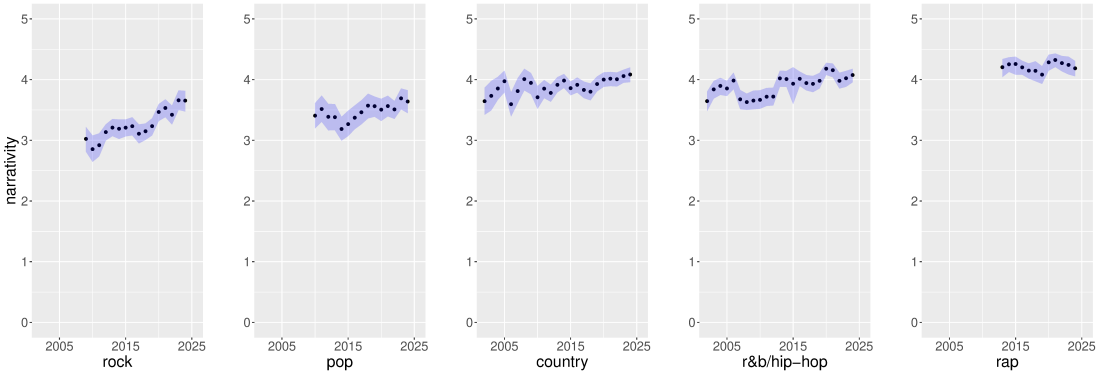

Figure 4: Narrativity over time, Billboard Year-End Hot Songs (by genre) with 95% confidence intervals.

6.3 Grammy nominations

Finally, we test the relationship between narrativity and prestige (measured through Grammy nominations) by comparing the average narrativity of Grammy-nominated songs to the narrativity of songs *from the same album*; this allows to control for factors such as the artist and year of release in influencing the nomination status of songs. We measure this for two sets of data: our set of manually annotated narrativity judgments (from the Grammy Song of the Year) and for automatically predicted narrativity scores for Grammy Best {Rock, Country, R&B, Rap} Song. We then carry out a paired t-test (pairing songs on the same album) to assess the significance of any difference in the mean narrativity between Grammy nominated-songs and non-nominated songs, applying a Bonferroni correction for the five hypothesis tests carried out.

For manually annotated Song of the Year tracks, nominated songs are 4.2% more narrative (3.15 vs. 3.02) but the difference is not significant when correcting for multiple hypothesis tests ($p = 0.050, n = 278$). For Grammy Best {Rock, Country, R&B, Rap} Songs, nominated songs are 3.1% more narrative (3.40 vs. 3.30) and significantly so ($p = 0.002, n = 768$), but a breakdown by genre (Table 3) reveals that this effect is due entirely to the Best Country Song category, which sees a 5.8% increase in narrativity between nominated songs and non-nominated songs from the same album (and significant at $\alpha = 0.05$).

Genre	Nominated	Non-nominated	↑	n	p
Country	3.61	3.40	5.8	259	0.002
R&B	3.03	2.98	1.6	247	0.426
Rock	3.13	3.08	1.5	162	0.502
Rap	4.25	4.18	1.6	100	0.261

Table 3: Narrativity by award category.

This finding is striking as scholarship around country music's prestige tends to emphasize a complicated conceptual terrain of "covert prestige" [13, p. 32] in tension with "constructed naturalness" [42, p. 3], where commercial success and mainstream prestige—the categories embodied in a Grammy Award—stand opposed to notions of "real country" [12; 22; 29]. Meanwhile, Jocelyn Neal has argued that specific "narrative paradigms" dominate country music composition [37],

but while she notes that the songwriters she discusses have won numerous awards, she does not make an explicit argument about the relationship between narrativity and industry prestige like our findings point to here. A focus on narrative might be a key way to rethink commercial prestige and the paradigms of composition that stretch across the genre, even amidst the competing models of prestige that inhere within country music.

7 Conclusion

In defining a computational model of narrativity in songs, this work charts the unequivocal rise of storytelling in contemporary popular music reflected in the Billboard top 100, finding an explanation for that rise in the increasing dominance of hip hop/rap in those charts and sustained by the renewed popularity of country in the 2020s. That hip hop, the genre most identified with lyric poetry, and not country, known for its long history of ballads, initially ushered in a new narrative moment in popular music, troubles the expected distinction between lyric and ballad. Pushing past these older divisions, hip hop requires a "lyrics theory," an understanding of how the most popular of song lyrics at the end of the 20th century and the first decades of the 21st fuse the poetic attention to rhythm and rhyme, and an intricate linguistic play with tropes and metaphors, to a propulsive narrative mode. This poetic combination of lyric and narrative also returns attention to classic studies of hip hop by Nelson George [14], Cheryl L. Keyes [27], and Tricia Rose [49], but with data to show that narrative storytelling isn't just a minor aspect of hip hop, but a major force that has shaped the recent history of popular music.[11] The two genres of hip hop and country have come to define the poles of our moment, with their cross pollination showing up in the music of Lil Nas X, Beyoncé and others; we might look to their shared focus on storytelling as a natural bridge between them.

Data and code to support this work, and enable other explorations of narrativity in song, can be found at `https://github.com/dbamman/song-narrativity`; this includes all manual annotations of 1,076 songs (linked to URLs on `www.azlyrics.com`); model narrativity predictions for the Billboard Hot 100 (1960-2024), Billboard genre subcharts, and Grammy nominees; and code to train and evaluate the models described in this paper.

Acknowledgments

The research reported in this article was supported by funding from the National Science Foundation (IIS-1942591), with computing resources provided by the Google Gemma Academic Program. This work benefitted from conversations with Colin Bazsali and helpful feedback from reviewers at CHR, for which we are grateful.

[11] Keyes [27], Robin D.G. Kelley [26], Rose [49], and others point to the history of "toasting" as the narrative form most closely aligned with hip hop. However, Joshi and Smith claim that "first-person toasts in the manner of Blowfy's 'Rapp Dirty' (1980) are, we found, relatively rare" [51, p. 148].

References

[1] Abrahams, Roger D and Foss, George. *Anglo-American Folksong Style*. Prentice-Hall, 1968.

[2] Akiba, Takuya, Sano, Shotaro, Yanase, Toshihiko, Ohta, Takeru, and Koyama, Masanori. "Optuna: A Next-Generation Hyperparameter Optimization Framework". In: *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2019, pp. 2623–2631.

[3] Alberhasky, Max and Durkee, Patrick K. "Songs tell a story: The arc of narrative for music". In: *PLOS ONE 19*, no. 5 (2024).

[4] Antoniak, Maria, Mire, Joel, Sap, Maarten, Ash, Elliott, and Piper, Andrew. "Where Do People Tell Stories Online? Story Detection Across Online Communities". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 7104–7130. DOI: 10.18653/v1/2024.acl-long.383.

[5] Bamman, David, Chang, Kent K, Lucy, Li, and Zhou, Naitian. "On classification with large language models in cultural analytics". In: *CHR* (2024).

[6] Bradley, Adam. *Book of rhymes: The poetics of hip hop*. Basic Civitas Books, 2009.

[7] Brysbaert, Marc, Warriner, Amy Beth, and Kuperman, Victor. "Concreteness ratings for 40 thousand generally known English word lemmas". In: *Behavior research methods 46* (2014), pp. 904–911.

[8] Culler, Jonathan. "Apostrophe". In: *Diacritics 7*, no. 4 (1977), pp. 59–69. ISSN: 03007162, 10806539.

[9] Culler, Jonathan. *Theory of the Lyric*. Cambridge, MA: Harvard University Press, 2015.

[10] Culler, Jonathan D. "Why Rhythm?" In: *Critical Rhythm: The Poetics of a Literary Life Form*, ed. by Ben Glaser and Jonathan D. Culler. New York: Fordham University Press, 2019.

[11] Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, and Toutanova, Kristina. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

[12] Edwards, Leigh H. "Country Music and Class". In: *The Oxford Handbook of Country Music*. Oxford University Press, July 2017. ISBN: 9780190248178. DOI: 10.1093/oxfordhb/9780190248178.013.19.

[13] Fox, Aaron A. *Real country: Music and language in working-class culture*. Duke University Press, 2004.

[14] George, Nelson. *Hip Hop America*. Viking Penguin, 1998.

[15] Greene, Roland, Cushman, Stephen, Cavanagh, Clare, Ramazani, Jahan, and Rouzer, Paul. *The Princeton encyclopedia of poetry and poetics*. Princeton University Press, 2012.

[16] Hampton, Timothy. *Bob Dylan's Poetics: How the Songs Work*. Zone Books, 2019.

[17] Hampton, Timothy. "Tangled Generation: Dylan, Kerouac, Petrarch, and the Poetics of Escape". In: *Critical Inquiry 39*, no. 4 (2013), pp. 703–731.

[18] Hankin, Charlie D. *Break and Flow: Hip Hop Poetics in the Americas*. University of Virginia Press, 2023.

[19] Harden, Alexander C. “Narrativizing recorded popular song”. In: *On popular music and its unruly entanglements* (2019), pp. 39–57.

[20] He, Pengcheng, Gao, Jianfeng, and Chen, Weizhu. “DeBERTaV3: Improving DeBERTa using electra-style pre-training with gradient-disentangled embedding sharing”. In: *arXiv preprint arXiv:2111.09543* (2021).

[21] Herman, David. *Basic elements of narrative*. John Wiley & Sons, 2009.

[22] Hughes, Charles. “Country Music and the Recording Industry”. In: *The Oxford Handbook of Country Music*. Oxford University Press, July 2017. ISBN: 9780190248178. DOI: 10.1093/oxfordhb/9780190248178.013.6.

[23] Hühn, Peter and Kiefer, Jens. *The narratological analysis of lyric poetry: studies in English poetry from the 16th to the 20th century*. Vol. 7. Walter de Gruyter, 2005.

[24] Ivanovs, Maksims, Kadikis, Roberts, and Ozols, Kaspars. “Perturbation-based methods for explaining deep neural networks: A survey”. In: *Pattern Recognition Letters 150* (2021), pp. 228–234. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2021.06.030.

[25] Jackson, Virginia. *Dickinson’s misery: A theory of lyric reading*. Princeton University Press, 2005.

[26] Kelley, Robin D. G. “Looking for the ‘Real’ Nigga”. In: *That’s the Joint: The Hip Hop Studies Reader*, ed. by Murray Forman and Mark Anthony Neal. New York: Routledge, 2004.

[27] Keyes, Cheryl L. *Rap music and street consciousness*. University of Illinois Press, 2002.

[28] Liu, Yinhan, Ott, Myle, Goyal, Naman, Du, Jingfei, Joshi, Mandar, Chen, Danqi, Levy, Omer, Lewis, Mike, Zettlemoyer, Luke, and Stoyanov, Veselin. “RoBERTa: A robustly optimized BERT pretraining approach”. In: *arXiv preprint arXiv:1907.11692* (2019).

[29] Lloyd, Richard. “The Sociology of Country Music”. In: *The Oxford Handbook of Country Music*. Oxford University Press, July 2017. ISBN: 9780190248178. DOI: 10.1093/oxfordhb/9780190248178.013.23.

[30] McLane, Maureen N. *Balladeering, minstrelsy, and the making of British romantic poetry*. Cambridge University Press, 2008.

[31] Mill, John Stuart. “Thoughts on Poetry and Its Varieties”. In: *Dissertations and Discussions: Political, Philosophical, and Historical*. Vol. 1. New York: Haskell, 1973, pp. 63–94.

[32] Mill, John Stuart. “What Is Poetry?” In: *Monthly Repository* (1833).

[33] Miller, Cristanne. “Hymn, the ‘Ballad Wild’ and Free Verse”. In: *Reading in Time: Emily Dickinson in the Nineteenth Century*. University of Massachusetts Press, 2012.

[34] Mire, Joel, Antoniak, Maria, Ash, Elliott, Piper, Andrew, and Sap, Maarten. “The Empirical Variability of Narrative Perceptions of Social Media Texts”. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, ed. by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 19940–19968. DOI: 10.18653/v1/2024.emnlp-main.1113.

[35] Modrow, Lena. *Wie Songs erzählen*. Frankfurt: Peter Lang, 2016.

[36] Moore, Allan F. “The persona-environment relation in recorded song”. In: *Rock Music, Routledge, London* (2017), pp. 275–294.

[37] Neal, Jocelyn R. “Narrative paradigms, musical signifiers, and form as function in country music”. In: *Music Theory Spectrum 29*, no. 1 (2007), pp. 41–72.

[38] Negus, Keith. “Narrative, interpretation, and the popular song”. In: *The Musical Quarterly 95*, no. 2-3 (2012), pp. 368–395.

[39] Neto, Pedro, Hartmann, Martin, Luck, Geoff, and Toiviainen, Petri. “An album is a story: Feature arcs in sequences of tracks”. In: *PLOS ONE 20*, no. 7 (July 2025), pp. 1–19.

[40] Neto, Pedro A.S.O., Hartmann, Martin, Luck, Geoff, and Toiviainen, Petri. “The algorithmic nature of song-sequencing: statistical regularities in music albums”. In: *Journal of New Music Research 52*, no. 5 (2023), pp. 410–424.

[41] Nicholls, David. “Narrative Theory as an Analytical Tool in the Study of Popular Music Texts”. In: *Music and Letters 88*, no. 2 (2007), pp. 297–315. ISSN: 00274224, 14774631.

[42] Pecknold, Diane. *The selling sound: The rise of the country music industry*. Duke University Press, 2007.

[43] Perry, Imani. *Prophets of the hood: Politics and poetics in hip hop*. Duke University Press, 2004.

[44] Pichler, Axel and Reiter, Nils. “From Concepts to Texts and Back: Operationalization as a Core Activity of Digital Humanities”. In: *Journal of Cultural Analytics 7*, no. 4 (2022).

[45] Piper, Andrew. “Think small: on literary modeling”. In: *PMLA 132*, no. 3 (2017), pp. 651–658.

[46] Piper, Andrew and Bagga, Sunyam. “NarraDetect: An annotated dataset for the task of narrative detection”. In: *Proceedings of the The 7th Workshop on Narrative Understanding*, ed. by Elizabeth Clark, Yash Kumar Lal, Snigdha Chaturvedi, Mohit Iyyer, Anneliese Brei, Ashutosh Modi, and Khyathi Raghavi Chandu. Albuquerque, New Mexico: Association for Computational Linguistics, May 2025, pp. 1–7. ISBN: 979-8-89176-247-3. DOI: 10.18653/v1/2025.wnu-1.1.

[47] Piper, Andrew and Bagga, Sunyam. “Toward a Data-Driven Theory of Narrativity”. In: *New Literary History 54*, no. 1 (2022), pp. 879–901.

[48] Piper, Andrew, Bagga, Sunyam, Monteiro, Laura, Yang, Andrew, Labrosse, Marie, and Liu, Yu Lu. “Detecting narrativity across long time scales”. In: *Proceedings of the Computational Humanities Research Conference* (2021).

[49] Rose, Tricia. *Black Noise: Rap Music and Black Culture in Contemporary America*. Hanover, NH: Wesleyan University Press, 1994.

[50] Simon, Julia. “Narrative Time in the Blues: Son House’s ”Death Letter” (1965)”. In: *American Music 31*, no. 1 (2013), pp. 50–72. ISSN: 07344392, 19452349.

[51] Smith, Macklin and Joshi, Aurko. *Rhymes in the Flow: How Rappers Flip the Beat (Kindle Edition)*. University of Michigan Press, 2020.

[52] Steg, Max, Slot, Karlo, and Pianzola, Federico. “Computational Detection of Narrativity: A Comparison Using Textual Features and Reader Response”. In: *Proceedings of the 6th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, ed. by Stefania Degaetano, Anna Kazantseva, Nils Reiter, and Stan Szpakowicz. Gyeongju, Republic of Korea: International Conference on Computational Linguistics, Oct. 2022, pp. 105–114.

[53] Stewart, Susan. “Notes on Distressed Genres”. In: *The Journal of American Folklore 104*, no. 411 (Winter 1991), pp. 5–31.

[54] Stimeling, Travis D. *The Oxford handbook of country music*. Oxford University Press, 2017.

[55] Stimeling, Travis D. “Narrative, vocal staging and masculinity in the ‘Outlaw’ country music of Waylon Jennings”. In: *Popular Music 32*, no. 3 (2013), pp. 343–358. ISSN: 02611430, 14740095.

[56] Ward, Andrew. *Popular song and narratology: Exploring the relationship between narrative theory and song lyrics through creative practice*. PhD thesis. Queensland University of Technology, 2019.

[57] Warner, Benjamin, Chaffin, Antoine, Clavié, Benjamin, Weller, Orion, Hallström, Oskar, Taghadouini, Said, Gallagher, Alexis, Biswas, Raja, Ladhak, Faisal, Aarsen, Tom, et al. “Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference”. In: *arXiv preprint arXiv:2412.13663* (2024).

A Author contributions

- David Bamman: Carried out computational experiments and analysis; data curation; wrote paper.
- Sabrina Baur: Operationalized the construct of narrativity for songs; designed annotation guidelines; annotated data; wrote paper.
- Mackenzie Hạnh Cramer: Operationalized the construct of narrativity for songs; designed annotation guidelines; annotated data; wrote paper.
- Anna Ho: Operationalized the construct of narrativity for songs; designed annotation guidelines; annotated data; wrote paper.
- Tom McEnaney: Contributed theoretical framing and analysis; wrote paper.

B Top Songs

What are the songs that show the highest narrativity over our period of analysis? We identify all songs from the Billboard Year-End Hot 100 from 1960–2024 and gather any available lyrics from `www.azlyrics.com` (a total of 5,745 songs). For each song, we use our best-performing model (RoBERTa) to predict its composite narrativity as the average of model predictions for Agent, Events, and World. Table 4 list the top ten songs by highest predicted narrativity.

Score	Date	Artist	Song
4.95	1973	Vicki Lawrence	The Night the Lights Went Out in Georgia
4.94	1967	Bobbie Gentry	Ode to Billie Joe
4.93	2003	Dierks Bentley	What Was I Thinkin'
4.93	2002	Kenny Chesney	The Good Stuff
4.92	1993	Ice Cube	It Was a Good Day
4.91	2022	Taylor Swift	All Too Well (Taylor's Version)
4.90	1972	Harry Chapin	Taxi
4.90	1968	Jeannie C. Riley	Harper Valley PTA
4.89	1989	Tone Loc	Wild Thing
4.89	2001	Craig David	Fill Me In

Table 4: Highest predicted songs among Billboard Hot 100 (1960-2024).

C Close Reading

Our computational model allows for what Bamman et al. 2024 [5] term "classification-assisted close reading"—identifying the salient texts within a broader collection for deeper analysis. Our top songs offer different engagements with narrativity, but across genres they vividly create worlds from romance and survival. The original 2012 version[12] of Taylor Swift's "All Too Well," overladen with flashbacks, scene cutting, and even metanarrative flourishes, culminates in the refrain's memory touchstone: "I remember it all too well." The opening line builds a world concisely through an "I" and "you"—the romantic dyad at the heart of pop romance—stepping into the threshold of the past: "I walked through the door with you." This spatialization of temporality (a doorway to the past) is also a particular kind of space: a domestic world, which "felt like home

[12] The version in our dataset is the 2021 re-recording ("Taylor's Version", 5 minutes and 29 seconds), which appears as #76 on the 2022 Billboard Hot 100 and has identical lyrics to the 2012 version.

somehow." The romantic memories that iterate across the song embed memory objects, none more potent than the lost scarf ("you've still got it in your drawer even now") from the opening verse that returns in the ultimate verse ("you keep my old scarf from that very first week"), tracing a narrative arc to a return that comments on the song's own theme of nostalgia and heartbreak.

The song's obsession with telling, with repeatedly narrating the inability to move on from this breakup, relates to the singer's ultimate power struggle to liberate herself from this past, and to make it a past where her ex-lover pines for her, where she retains control. Whereas at the start of the final verse the singer was stuck—"Time won't fly, it's like I'm paralyzed by it"—by the end of that verse the tables seem to be turned. The scarf marks her victory—"And it smells like me / You can't get rid of it"—and seems to put her ex in a place where he longs for her. The closing refrain ("Wind in my hair, you were there, you remember it all / Down the stairs, you were there, you remember it all") nearly closes the song with him stuck in the past, remorseful and desiring her. And yet, in the final line, the singer seems to join that same position: "I remember it all." Is this a generous attempt at reconciliation? Or is it a structural conflict for the ego, the insistence that the "I" take the spotlight again in the end, even if it seems to turn her backwards as the song concludes?

The high narrativity of this song derives from its agonic romance (the struggle between the agents "I" and "you"), it's detailed world building ("your drawer," "getting lost upstate," "autumn leaves," "that little town street," "a twin-sized bed"; "the refrigerator light"), its evolving motifs ("you almost ran the red"; "your cheeks were turning red"; "I can picture it"; "Photo album on the counter"), and an insistence on temporality and tense shifts ("we're singing in the car"; "there we are again on that little town street"; "you used to be a little kid with glasses"; "there we are again in the middle of the night") that culminate in the song's refrain "you / I remember it all" linked to that sense of temporal stasis ("time won't fly") and the tension between looking back, feeling the past as present, or moving on. We don't need Faulkner or Genette to tell us narrative isn't necessarily about forward progression. Our model recognizes that stories, from national traumas to personal romances, can narrate the inability to move on, a sentiment embodied in one of the world's most famous singers revising her own heartbreak anthem in a sonic tumble of fading repetitions.

Ice Cube's "It Was A Good Day" joins the pantheon of attempts to narrate a life by telling the story of a day in the life. The story's motivation is romantic, just like Swift's, but there's only teleology here, as Cube gets a call from "a girl" in the first verse and picks her up in the sixth verse, the day concluding after their romantic connection and a different version of Swift's emphasis on the "I": "Even saw the lights of the Goodyear Blimp / And it read 'Ice Cube's A Pimp.'" A more public affirmation than in Swift, the narrative world recognizes the romantic aim of the song, its lights contrasting with what the singer doesn't see: a cop car's "high beams," "No helicopter looking for the murder." These other forces typically stand in the way of this singer's journey. Different from Swift's song, Ice Cube's antagonism lies with the police and rivals whose absence on this day ("not a jacker in sight"; "saw the police and they rolled right past me") allows him to connect and share with his lover ("I had the brew, she had the chronic"). This isn't a story about a romantic struggle. It's a narrative of survival in a racist state.

D Annotation Guidelines

Adapted from the annotation codebook of Piper and Bagga [47].

Task: After reading a song's lyrics in its entirety, annotators rate their agreement with the following statements.

D.1 Statement 1: This passage foregrounds lived experience of particular agents

"Are there a limited number of agents who are clearly foregrounded and experiencing something in this passage? I.e. is there a protagonist or two primary characters that run through all the actions? The more centralized and consistent one or more agents are and the more coherent their identities, then the stronger"agency" is as a quality. Jumping from one entity to the next would constitute low agency, even if each of those agents does something. How focalized is the passage around a central figure(s)?" - Bagga & Piper.

Qualities to consider:

- "Lived experience:" How specific are the protagonists' actions and experiences? Are they performing key actions (ie. a romantic picnic, a quest)? Are they feeling particular emotions (ie. regret, anger, sadness, euphoria)? Are they explaining the background that led to those actions and/or emotions?

- "Particular agents:" How specific are protagonists' identities and personalities? Do we know who they are? Do we know about their past, their families, their professions? Could they not be replaced with people of any character, class, gender, nationality, race, appearance, location etc.?

  - If both those qualities are met, mark 5/high agreement.
    - Example marked 5: Cat's Cradle.
      - A father growing older (PA) no longer has a relationship with his son, whom he had neglected due to work (LE).

  - If only one of those qualities are met, mark 2-4 for middle/low agreement, depending on how specific the met element is.
    - Example marked 2: Eight Days a Week
      - The narrator cares about a girl and thinks about her everyday. This is missing PA (we have no details about who the narrator is) but has some LE of them being in love.

    - Example marked 4: august
      - The narrator reminisces about a bygone relationship (LE). We don't know who the narrator is, but the level of detail in the described memories allows us to get a sense of their priorities in those moments.

  - If neither are met or are only barely met, mark 1/low agreement.
    - Example marked 1: Fame
      - We know nothing about the narrator except their opinions on fame. There is no protagonist.

Potential ambiguities:

- Lived experience can help narrow down the particular agent. Someone reacting abnormally to an event, engaging in drastic activity, or feeling complicated/contradictory emotions can also give insight to their character.

  - Example marked 5: Norwegian Wood.

        * The narrator is disappointed in love after a girl invites him back to her place. In response, he lights a fire. This specific action is evocative of an ending scene that shows the particularity of the agent in question.

- Protagonist is not always the narrator and vice versa. If the song has multiple protagonists, do most of them fit within the above qualities? Rate accordingly.

D.2 Statement 2: The passage is organized around sequences of events that occur over time

"How clearly do you see a sequence of events in your passage? I.e. can you put"then" into the sentences easily (then this, then this)? They may be temporally out of order (I tell what happened last first) or there may be gaps between their happening (I walk into a restaurant and then go home), and there may be simultaneity (something happens at the same time as the thing before it), but there is an underlying temporal relationship between the events that are mentioned in the passage (i.e. they connect to each other in an experiential way)." - Bagga & Piper.

Qualities to consider:

- "Sequences of events:" The sequence of events in particular is distinct from the condition -- a condition describes a state of being whereas the events recount what led to a particular state and the phenomena of the state. For example, "I am sad" could be seen as an event, but it does not qualify as a high agreement S2 unless it is attached to a series of other events.

- "Over time:" Not all events may be described in equal detail, but there should be enough specificity to place all of them on a timeline as well as logical/causal/situational links, ie. X happened so I am sad and I did Y.
    - If qualities are met in detail, mark 5/high agreement.
        * Example marked 5: Cat's Cradle.
            · The deterioration of the father-son relationship can be traced to chronologically-depicted events in each verse.

    - If qualities are met generally, mark 2-4 for middle/low agreement, depending on how specific the met element is.
        * Example marked 4: Beat it
            · In dispensing advice for a hunted individual to "beat it" the narrator also discloses past and present happenings. They also foreshadow what will happen if his advice is not heeded. The general sentiment of the events stay the same but cohere well.

    - If qualities only barely met, mark 1/low agreement.
        * Example marked 1: Fame
            · This doesn't describe events in sequence but a phenomenon.

Potential ambiguities:

- Future events and degree of intentionality
    - It can be difficult to tell which events count toward the "narrative timeline" if they're being forecasted. Place them according to the intentionality of the narrator (ie. Making a wish/hope should not be counted as an event versus making a plan or stating "I will do X.")

D.3 Statement 3: The passage creates a world that I can see or feel.

"Think about this as"inhabitability" -- can you inhabit what is happening in your "mind's eye"? Can it be experienced? Also key here is the unity of the world -- the actions all make sense as part of a unified coherent space, even if this is a fantastic space. Things are happening together. World making isn't exclusively about description -- very underdescribed worlds can be very concrete in terms of their inhabitability." - Bagga & Piper.

Qualities to consider:

- Inclusion of sensory elements: Can we see/feel/smell/touch the world being presented? Are there physical elements, objects, furnishings or geographical features with which we can anchor our understanding? Is this world populated with other people besides the protagonist? The scale of the 'world' may be as vast as the western hemisphere or as small as a bedroom, but the amount/specificity of sensory details will determine agreement.

- "Indexing:" 'real world' elements such as current events, place names, popular figures, trends or vernacular can provide additional dimension to the passages' 'worldliness' as a situating short-cut.

  - If at least one of these qualities is met in detail, mark 5/high agreement.
    - Example marked 5: Cat's Cradle.
      - Though the father-son relationship is front and center in this narrative, we are ushered into this world via references to traveling, careers, bonding time and the ever-repeating nursery rhymes.

  - If at least one of these qualities are briefly met, mark 2-4 for middle/low agreement, depending on how specific the met element is.
    - Example marked 2: I Still Haven't Found
      - The landscapes mentioned in this song (ie "highest mountain") are primarily to highlight the drastic actions the protagonist has taken rather than to create a sense of the setting.

    - Example marked 4: Beautiful
      - In portraying body image troubles, the narrative alludes to a world where the protagonist is shamed into feeling insecure and "delirious."

  - If neither are met or are only barely met, mark 1/low agreement.
    - Example marked 1: Fame
      - The only setting element mentioned is a limo, briefly.

Potential ambiguities:

- Unrealism: descriptions may feel less genuine or visceral if interpreted as extended metaphors or allegories (which they sometimes are!). They may also be cliched, stereotypical, or hyperbolic, all of which could affect the annotator's ability to truly "inhabit" what is being portrayed. However, we should prioritize the level of detail supplied when annotating over the realism of the events/world.

  - Example marked high agreement: Viva La Vida
    - The narrator recounts his fall from grace with a Roman Empire-esque backdrop. Though the setting seems highly stylised, the level of detail and cohesion qualifies it for a higher agreement rating.

D.4 Additional considerations:

- Overlap between S1's Lived Experiences and S2
  - Though both elements involve narrative events, S1 is focused on the protagonist's experience of those events while S2 is interested in their sequencing (ie. how cohesively the events fit together). They should be considered as separate elements that can drastically differ.
    - Example marked 4 for S1, 2 for S2: Hungry Heart
      - The narrator describes leaving a wife and kids as well as engaging in a failed relationship in Kingstown. Though his lived experience is palpable and detailed, it's unclear how one action (driving away from a wife and kids) relates to the other (falling in love then ending a relationship), or if the two are even connected. Their ordering is also unclear.
    - Example marked 3 for both S1 and S2: They Said You Needed Me
      - Example of I-statements combined with action verbs contributing to both lived experiences and event sequences. It demonstrates how a passage's key features can be attributed to multiple categories.
- Level of inference
  - Many works evoke past agents or events only briefly and we must rely on inference in order to make categorical judgements. The level of inference necessary (high vs. low) can depend on the amount and specificity of existing details. In Norwegian Wood, for instance, the main skeletal elements required for high-agreement annotation are all present despite a dearth of description. Songs requiring high levels of inference may be judged based on how much is being implied.
- Cross-category influence
  - Though the statements' narrative qualities may inevitably influence one another at times (ie. S3's lack of world-sense is affected by S1's narratorial ambivalence), we are encouraged to consider each statement as its own element and consider them independently from one another. The categories may have similar or vastly different agreement levels.

E LLM prompt

The following is the prompt given to LLMs (GPT-4.1, Gemini 2.5 Pro, and Claude Opus 4).

Task: After reading a song's lyrics in its entirety, rate your agreement with the following statements on a Likert Scale (1, 2, 3, 4 or 5), with 1 denoting low agreement and 5 denoting high agreement.

Q1: This passage foregrounds lived experience of particular agents. Are there a limited number of agents who are clearly foregrounded and experiencing something in this passage? I.e. is there a protagonist or two primary characters that run through all the actions? The more centralized and consistent one or more agents are and the more coherent their identities, then the stronger "agency" is as a quality. Jumping from one entity to the next would constitute low agency, even if each of those agents does something. How focalized is the passage around a central figure(s)?

Q2: The passage is organized around sequences of events that occur over time. How clearly do you see a sequence of events in your passage? I.e. can you put "then" into the sentences easily (then this, then this)? They may be temporally out of order (I tell what happened last first) or there may be gaps between their happening (I walk into a restaurant and then go home), and there may be simultaneity (something happens at the same time as the thing before it), but there is an underlying temporal relationship between the events that are mentioned in the passage (i.e. they connect to each other in an experiential way).

Q3: The passage creates a world that I can see or feel. Think about this as "inhabitability" – can you inhabit what is happening in your "mind's eye"? Can it be experienced? Also key here is the unity of the world – the actions all make sense as part of a unified coherent space, even if this is a fantastic space. Things are happening together. World making isn't exclusively about description – very underdescribed worlds can be very concrete in terms of their inhabitability.

Format your response as a json file: {"Q1": N, "Q2": N, "Q3": N}. Here are sample inputs/outputs:

Input: (song 1)

Output: {"Q1": 5, "Q2": 5, "Q3": 5}

Input: (song 2)

Output: {"Q1": 5, "Q2": 2, "Q3": 2}

Input: (song 3)

Output: {"Q1": 1, "Q2": 1, "Q3": 1}

Input: (target song)

Output:

F Model performance by narrativity axis

Model	Agent ρ	Event ρ	World ρ
RoBERTa	0.682 [0.587-0.761]	0.786 [0.733-0.830]	0.816 [0.761-0.858]
DeBERTa-v3	0.657 [0.568-0.733]	0.792 [0.741-0.834]	0.802 [0.747-0.847]
BERT	0.614 [0.516-0.697]	0.760 [0.698-0.811]	0.812 [0.755-0.854]
ModernBERT	0.584 [0.474-0.675]	0.743 [0.673-0.801]	0.745 [0.675-0.802]
BOW/POS/an./conc/1pp	0.669 [0.571-0.749]	0.721 [0.645-0.784]	0.749 [0.684-0.802]
`Gemini 2.5 Pro`	0.213 [0.047-0.361]	0.566 [0.463-0.654]	0.603 [0.509-0.685]
`GPT 4.1`	0.336 [0.199-0.462]	0.600 [0.496-0.690]	0.456 [0.332-0.566]
`Claude Opus 4`	0.281 [0.139-0.411]	0.618 [0.514-0.704]	0.515 [0.403-0.616]
POS/an./conc/1pp	0.353 [0.227-0.469]	0.542 [0.441-0.631]	0.492 [0.387-0.589]
Minimal narrative feats.	0.273 [0.145-0.392]	0.474 [0.366-0.572]	0.368 [0.248-0.480]
1pp	0.090 [-0.046-0.227]	-0.091 [-0.224-0.043]	0.213 [0.077-0.342]

Table 5: Model performance along each individual axis, along with 95% bootstrap confidence intervals; models are ranked from best to worst based on composite narrativity (as in table 1).

Table 5 illustrates the performance of all models along each individual axis (ranked according to their overall performance on the composite narrativity score). We can see that the agent axis is the most difficult for all models, and especially so for prompted-based LLMs.

G Interpretability

In order to understand what the best-performing RoBERTa model is learning about narrativity, we draw on perturbation-based methods [24] to interrogate which words in the input are most critical for the predictions being made. Given a textual sequence $x = \{x_1, \ldots x_n\}$ corresponding to the lyrics of a song, we replace each single token x_i in turn with RoBERTa's mask symbol (one token at a time) and generate a prediction from that perturbation. We treat the tokens corresponding to the 10 lowest-scoring perturbations (i.e., the tokens whose removal leads to the biggest reductions in narrativity) as the most significant *local* explanations (the leading explanations for the predictions made for a single song). We aggregate these local, song-level explanations into global model explanations by counting the fraction of times a given word type appears in a local explanation among all of its occurrences.

The most frequent terms in local explanations for the agent task include not only a focus on the first-person (*I*) and *mirror* reflecting back on that first person, but also a focus on inebriation (*drunk, wine, wasted, drank*). The most common explanations for the event task include not only a focus on the past tense identified by Piper and Bagga (*said, planned, stole*), but also explicit temporal indicators (*tonight, yesterday, evening, tomorrow*). The world-building task includes explanations in specific places (*downtown, hall, town, mall, street, train, city*) and things *(coffee, wine)*.

H Robustness

To test the robustness of these results, we probe two adversarial scenarios: first, is it possible that a BERT-based model (or any other learned model) has learned a bias in its predictions that invents a trend over time that is not present in the original annotations? Figure 5 rejects this by illustrating the average narrativity over time only for human-annotated songs (i.e., not model predictions); while the confidence intervals are much wider (corresponding to the smaller amount of data annotated per year), we see the same increasing trend.

agent		events		world	
mirror	0.538	tonight	0.792	downtown	0.710
I	0.531	telephone	0.704	hall	0.629
drunk	0.479	said	0.632	town	0.625
wine	0.448	yesterday	0.629	mall	0.608
affair	0.442	planned	0.618	coffee	0.603
wasted	0.424	evening	0.611	street	0.597
conversation	0.386	tomorrow	0.606	train	0.596
car	0.361	home	0.594	city	0.578
funny	0.356	stole	0.575	wine	0.512
drank	0.343	were	0.573	moonlight	0.507

Table 6: Top global explanations for each narrativity task. 79.2% of occurrences of *tonight* appear as local explanations.

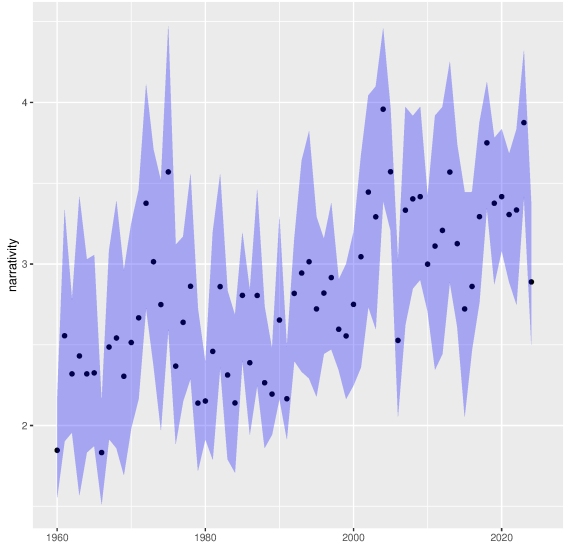

Figure 5: Narrativity over time, measured only using manual annotations.

Second, is it possible that judgments of narrativity are influenced by annotators' familiarity with the songs being judged? A credible alternative is that any human judge may see songs they are familiar with as more narrative than songs that are not, simply as a function of having a deeper contextual understanding of those songs developed through repeated listening. Annotators could also in principle draw on musical and visual information (in the context of music videos) in implicitly informing their judgments. We test this by asking annotators to rate their familiarity with a subset of songs they annotated, and only plotting the narrativity of songs they are *not* familiar with. Fig. 6 illustrates that this subset displays the same increase over time as the predicted measures of narrativity of full Billboard charts.

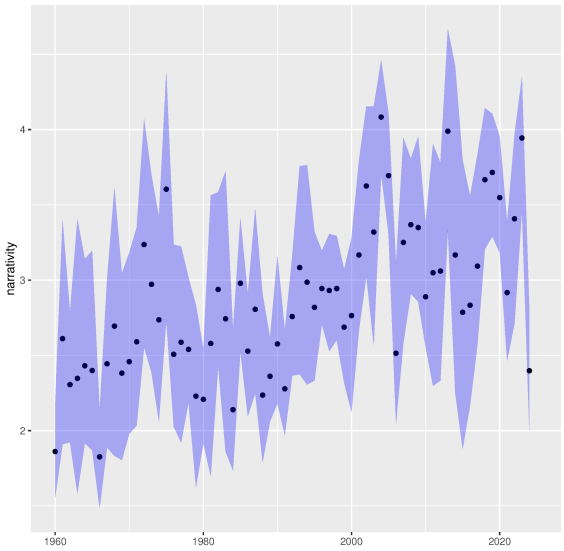

Figure 6: Narrativity over time, average annotation for songs *unknown* to annotators. Increasing narrativity over time is not explained by annotator familiarity.