# Can Transformers Hear Genre Through Lyrics Alone?

A Comparative Study of Encoder-Only, Decoder-Only, and Encoder-Decoder Architectures with Interpretability Analysis

Junho Hong
*STAT 359 — Final Project*
hjunho@u.northwestern.edu

March 2026

**Abstract**

We investigate how well transformer models can classify music genre from song lyrics alone, comparing three architectural paradigms—encoder-only (RoBERTa-base), decoder-only (GPT-2), and encoder-decoder (T5-small)—all fine-tuned with Low-Rank Adaptation (LoRA). On a balanced five-genre dataset (Rap, Pop, Rock, Country, R&B; $n \approx 10{,}000$), RoBERTa achieves the highest macro F1 of 0.611, followed by GPT-2 (0.591) and T5 (0.463), though these rankings are confounded by differences in base model size and classification head capacity that we analyze in detail. All approaches plateau well below perfect accuracy, prompting the central question of this work: *why?* We answer with four converging interpretability analyses on RoBERTa: (1) layer probing shows genre information saturates by layer 7 of 12, indicating a representational ceiling; (2) contrastive Integrated Gradients reveal that model confusion tracks genuine lexical ambiguity between similar genres; (3) embedding visualizations show Pop, Rock, and R&B representations are heavily interleaved; and (4) an independent genre similarity analysis using Sentence-BERT finds that Pop–Rock centroid cosine similarity reaches 0.979. External validation from zero-shot Claude Opus ($\text{F1} = 0.636$) corroborates the ceiling. These findings establish that lyrics-only genre classification is inherently bounded at approximately $\text{F1} \approx 0.65$—a data-driven ceiling requiring multimodal features to overcome.

## 1 Introduction

Music genre classification is a foundational task in music information retrieval (MIR) with direct applications in streaming recommendation, content moderation, and musicological research. While genre has traditionally been defined by auditory characteristics—tempo, instrumentation, harmonic structure—song lyrics provide a complementary textual signal that captures thematic and linguistic patterns unique to each genre.

The emergence of large pretrained language models has opened new avenues for text classification. However, genre classification from lyrics presents a distinctive challenge: genres are socially constructed categories with fuzzy boundaries, and many songs intentionally blend genres. A Pop ballad may share the lyrical themes of a Country song, while an R&B track may employ vocabulary similar to Rap. This raises two questions: *how well can transformers classify genre from lyrics alone?* And when performance plateaus, is the bottleneck the model or the data?

We address these questions in two stages. First, we conduct a controlled comparison of three transformer paradigms—RoBERTa (encoder-only), GPT-2 (decoder-only), and T5 (encoder-decoder)—using LoRA adapters [Hu et al., 2022] with matched adapter budgets but architecture-specific total trainable counts (0.59–1.18M). All three models plateau at similar F1 levels (0.46–0.61), well below perfect accuracy. Second, rather than chasing incremental improvements,

we investigate *why* the plateau exists. Through interpretability analyses on RoBERTa—layer probing, contrastive Integrated Gradients, embedding visualization—combined with an external genre similarity analysis and zero-shot LLM baselines, we show that the ceiling is *data-driven* (genres genuinely overlap in lyric content) rather than model-driven.

Our contributions are:

- A controlled comparison of three transformer paradigms for lyrics-based genre classification under LoRA fine-tuning, with transparent reporting of confounds (head size, base model size) that complicate direct architectural comparison.
- Interpretability evidence explaining *why* performance plateaus: layer probing reveals representational saturation at mid-depth; contrastive attribution shows errors arise from genuine lyrical ambiguity; and embedding visualizations confirm confused genres are interleaved in representation space.
- Converging external evidence that the text-only ceiling is approximately $F1 \approx 0.65$: genre centroid similarity of 0.979 (Pop–Rock), heavily overlapping within/between-genre similarity distributions, and a Claude Opus zero-shot baseline of only 0.636.

## 2 Related Work

**Music Genre Classification.** Early approaches to genre classification relied on handcrafted audio features such as Mel-frequency cepstral coefficients (MFCCs), spectral centroid, and tempo [Tzanetakis and Cook, 2002]. Lyrics-based methods initially used bag-of-words and TF-IDF representations with traditional classifiers [Fell and Sporleder, 2014]. Tsaptsinos [2017] applied hierarchical attention networks to lyrics, demonstrating that deep learning could surpass hand-engineered features for this task. More recent work has combined audio and text modalities [Oramas et al., 2018], though lyrics-only classification remains an active area due to its accessibility and interpretability.

**Transformers for Text Classification.** BERT [Devlin et al., 2019] and its variants have become the dominant paradigm for text classification. RoBERTa [Liu et al., 2019] improved upon BERT through extended pretraining and dynamic masking. Decoder-only models like GPT-2 [Radford et al., 2019] have been adapted for classification via sequence-level pooling, while encoder-decoder models like T5 [Raffel et al., 2020] reframe classification as text-to-text generation. Direct comparisons across all three paradigms for the same classification task remain relatively scarce.

**Parameter-Efficient Fine-Tuning.** Full fine-tuning of large language models is computationally expensive and prone to overfitting on small datasets. Low-Rank Adaptation (LoRA) [Hu et al., 2022] addresses this by injecting trainable low-rank matrices into attention layers while freezing pretrained weights, enabling roughly matched trainable parameter budgets across models of different sizes.

**Probing and Interpretability.** Probing classifiers [Conneau et al., 2018] have become the standard tool for analyzing what information is encoded at each layer of a pretrained model. Tenney et al. [2019] showed that BERT's layers recapitulate the classical NLP pipeline, with syntactic information peaking in middle layers and semantic information in later layers. For token-level attribution, Integrated Gradients [Sundararajan et al., 2017] provides axiomatic guarantees of completeness and sensitivity.

**Zero-Shot LLM Classification.** The in-context learning paradigm introduced by GPT-3 [Brown et al., 2020] enables large language models to perform classification without task-specific training. We use zero-shot LLM evaluation as a reference point that provides an upper bound on the information accessible from lyrics plus world knowledge.

# 3 Dataset

## 3.1 Data Collection and Preprocessing

We source song lyrics from the `genius-song-lyrics` dataset on Hugging Face [Dizon, 2024], which contains 2.76 million songs with lyrics and genre metadata scraped from Genius.com. We retain five genres: **Rap/Hip-Hop**, **Pop**, **Rock**, **Country**, and **R&B**. The preprocessing pipeline:

1. **Language filtering:** Retain only English-language songs, requiring agreement between CLD3 and fastText detectors.
2. **Stratified sampling:** Sample 2,000 songs per genre (2,500 for Pop to offset higher noise) for the comparison experiment. A larger set of 5,000 per genre (6,000 for Pop) is used for interpretability analyses. All sampling uses seed 42.
3. **Lyric cleaning:** Remove structural annotations (e.g., `[Verse]`, `[Chorus]`) via regex, collapse excessive whitespace, and truncate to 2,000 characters.
4. **Minimum length filter:** Discard songs with fewer than 20 words after cleaning.
5. **Stratified splitting:** 80/10/10 train/validation/test split with stratification by label. Validation and test sets are further balanced to equal counts per class.

The resulting comparison dataset contains approximately 8,000 training, 1,000 validation, and 1,000 test samples.

## 3.2 Exploratory Data Analysis

Figure 1 shows substantial variation in lyric length across genres. Rap lyrics are the longest (mean 333 words), while Rock are the shortest (mean 196 words). This length disparity could serve as a confounding feature; however, tokenizer truncation at 512 tokens bounds the effective input length.

TF-IDF analysis (Figure 2) reveals that each genre possesses a distinctive vocabulary: Rap is characterized by explicit language and slang; Rock by themes of darkness and pain; Country by domestic and rural imagery (*old*, *home*, *town*); R&B by terms of endearment. Notably, Pop has the weakest distinctive terms—its most characteristic words (*away*, *day*, *night*) are generic and shared with other genres.

Vocabulary overlap analysis (Figure 3a) confirms that Rap has the most distinctive lexicon, while Pop, Rock, and Country share substantial vocabulary. These EDA findings foreshadow the classification results: genres with more distinctive linguistic signatures are easier to classify.

# 4 Methods

## 4.1 Model Architectures

We compare three pretrained transformer architectures:

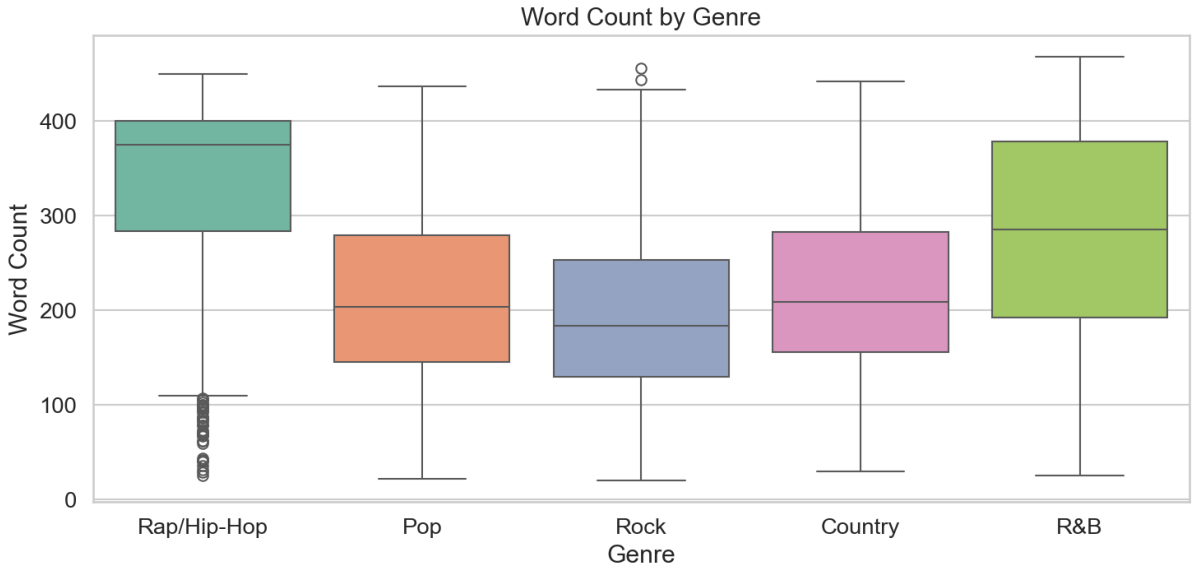



Figure 1: Word count distributions by genre. Rap lyrics are substantially longer (median 375 words) than Rock (184) or Pop (204), reflecting the density of rap verses. All models receive the same 512-token truncated inputs.

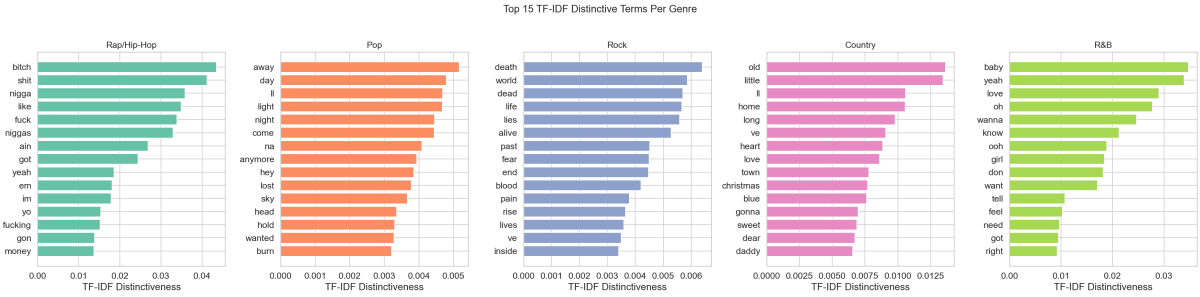



Figure 2: Top 15 TF-IDF distinctive terms per genre, computed as the difference between a genre's mean TF-IDF vector and the complement's mean. Pop has the weakest distinctive terms, foreshadowing its role as the most-confused class.

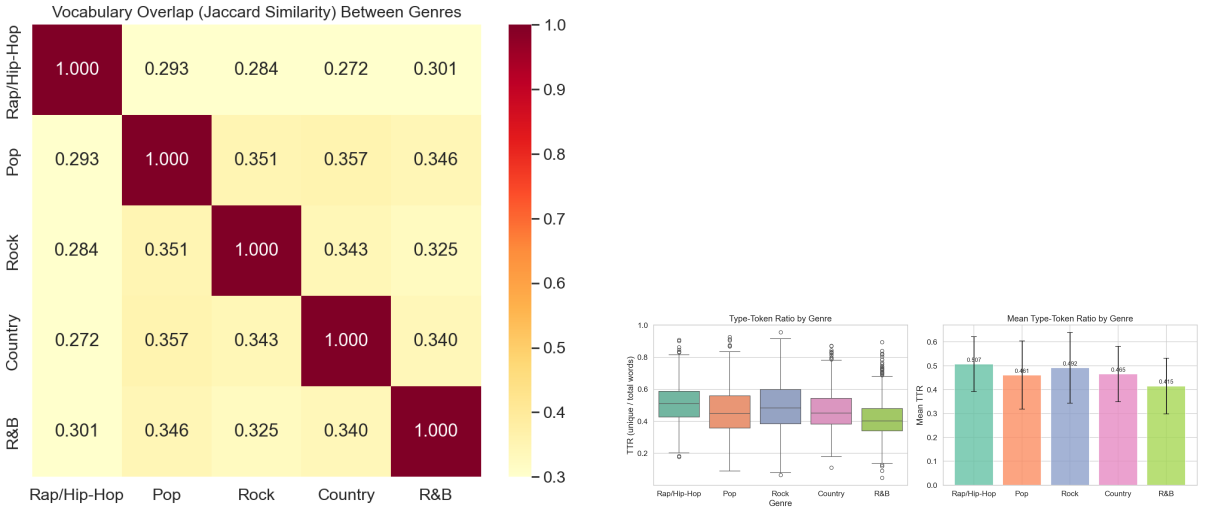
**RoBERTa (Encoder-Only).** `roberta-base` (125M parameters) with a two-layer classification head on the `[CLS]` token [Liu et al., 2019]. Bidirectional self-attention allows each token to attend to the full context.

**GPT-2 (Decoder-Only).** `gpt2` (124M parameters) with a classification head on the last non-padding token [Radford et al., 2019]. Causal (left-to-right) attention forces sequential representation building; padding token is set to `eos_token`.

**T5 (Encoder-Decoder).** `t5-small` (60M parameters) in text-to-text mode [Raffel et al., 2020]: input is `"classify genre: <lyrics>"`, target is the genre string. Predictions are decoded via greedy generation (`max_new_tokens=8`) and mapped back to integer labels.

### 4.2 Parameter-Efficient Fine-Tuning with LoRA

All models are fine-tuned using LoRA [Hu et al., 2022], which injects trainable rank-$r$ decomposition matrices $\Delta W = BA$ into selected attention projections while freezing all pretrained parameters. Table 1 summarizes the configuration.




(a) Jaccard vocabulary overlap.

(b) Type-token ratio by genre.

Figure 3: (a) Pairwise Jaccard similarity of unique word sets per genre (stopwords removed). Rap shares the least vocabulary with other genres. (b) Vocabulary richness: Rock and Country exhibit higher TTR (more diverse word choices per song).

Table 1: LoRA configuration ($\alpha = 32$, dropout $= 0.1$). Adapter parameters are matched (590K); total trainable counts differ due to classification heads.

| Model | Base Size | Targets | Rank | Head | Total Trainable |
|---|---|---|---|---|---|
| RoBERTa | 125M | query, value | 16 | 594K | 1.18M |
| GPT-2 | 124M | c_attn | 16 | 3.8K | 0.59M |
| T5 | 60M | q, v (enc.+dec.) | 16 | 0 | 0.59M |
| RoBERTa (interp.) | 125M | query, key, value | 16 | 594K | 1.48M |

While the LoRA adapter parameters are identical across the comparison models (590K), total trainable counts differ: RoBERTa's two-layer classification head (768→768→5) adds 594K parameters, GPT-2's single linear head adds only 3.8K, and T5 has no separate head (text-to-text reuses the frozen LM head). This asymmetry complicates direct comparison and is discussed in Section 7.4.

### 4.3 Training Procedure

**Optimization.** All models are trained with AdamW [Loshchilov and Hutter, 2019] and a cosine learning rate schedule with linear warmup. Gradient norms are clipped at 1.0.

**Early Stopping.** We monitor validation macro F1 and halt if no improvement is observed for 3 consecutive epochs (comparison) or 4 epochs (interpretability model). The checkpoint with the highest validation F1 is retained.

**Interpretability Model.** For interpretability analyses (Section 6), we train a separate RoBERTa on the larger dataset (5,000 samples/genre) with expanded LoRA targets (query/key/value), a lower learning rate ($10^{-4}$), gradient accumulation (effective batch size 64), FP16 mixed precision, and random contiguous word cropping as data augmentation (50–100% of words for songs >100 words). This model achieves F1 = 0.608 on the Phase 2 test set and 0.605 on the original test

set—comparable to the comparison RoBERTa (0.611)—confirming it is a representative subject for interpretability.

Table 2: Hyperparameter settings.

| **Hyperparameter** | **Comparison** | **Interpretability** |
|---|---|---|
| Max sequence length | 512 | 512 |
| Batch size | 64 | 32 |
| Learning rate | $2 \times 10^{-4}$ | $1 \times 10^{-4}$ |
| Weight decay | 0.01 | 0.01 |
| Max epochs | 10 | 15 |
| Warmup ratio | 0.10 | 0.06 |
| Gradient accumulation | 1 | 2 |
| Mixed precision (FP16) | No | Yes |
| Augmentation | No | Yes |
| Samples per genre | 2,000 | 5,000 |

### 4.4 Zero-Shot LLM Baselines

To contextualize fine-tuned model performance, we evaluate two Claude models [Anthropic, 2025b,a] in a zero-shot setting following the in-context learning paradigm of Brown et al. [2020]: **Claude Haiku** (smaller, faster) and **Claude Opus** (larger, more capable). Each model receives a system prompt constraining output to a single genre label, and the input is the raw lyric text (truncated to 3,000 characters). We evaluate on 50 randomly sampled test songs per genre (250 total). Unlike our fine-tuned models, these LLMs have access to world knowledge about artists, musical styles, and cultural context, providing a reference point for the information available beyond lyrics alone. We note that the 250-sample evaluation introduces higher variance than the full 1,000-sample test set, and treat LLM results as illustrative rather than definitive.

### 4.5 Interpretability Methods

All interpretability analyses are conducted on the RoBERTa model trained on the larger dataset.

**Genre Similarity Analysis.** To quantify task difficulty independently of our trained models, we encode all lyrics using a sentence-transformer (`all-MiniLM-L6-v2`; Reimers and Gurevych, 2019) and compute pairwise centroid cosine similarities between genres. Higher inter-genre similarity implies harder classification. We also compare within-genre and between-genre pairwise similarity distributions to assess the degree of class overlap.

**Layer Probing.** Following Conneau et al. [2018] and Tenney et al. [2019], we extract the `[CLS]` embedding from each of the 13 layers (layer 0 = static embeddings; layers 1–12 = transformer outputs) and train a multinomial logistic regression classifier (L-BFGS solver, $C = 1.0$, max 1,000 iterations) on the training set. Test-set macro F1 at each layer quantifies linearly accessible genre information at that depth. We note that probing accuracy is a lower bound on the information present, since it reflects only what a linear classifier can extract.

**Embedding Visualization.** We extract `[CLS]` embeddings from the final transformer layer and project them to 2D using t-SNE [van der Maaten and Hinton, 2008] (perplexity 30, 1,000 iterations) and UMAP [McInnes et al., 2018] ($k = 15$, min_dist = 0.1). These projections are qualitative; we avoid over-interpreting cluster boundaries given the stochastic nature of both methods.

Table 3: Test-set performance. Fine-tuned models are evaluated on the full 1,000-sample test set. LLM baselines use a 250-sample subset (50 per genre). All scores are single-run point estimates; see Section 7.4 for discussion of variance.

| Model | Type | Accuracy | F1 Macro | F1 Weighted |
|---|---|---|---|---|
| RoBERTa | Fine-tuned | 0.613 | **0.611** | 0.611 |
| GPT-2 | Fine-tuned | 0.604 | 0.591 | 0.591 |
| T5 | Fine-tuned | 0.454 | 0.463 | 0.463 |
| Claude Haiku | Zero-shot LLM | 0.596 | 0.584 | 0.587 |
| Claude Opus | Zero-shot LLM | **0.649** | **0.636** | 0.639 |
| *Random* | *Baseline* | *0.200* | *0.200* | *0.200* |

**Contrastive Integrated Gradients.** For the top confused genre pairs (identified from the confusion matrix), we apply Integrated Gradients [Sundararajan et al., 2017] (implemented via Captum) with a *contrastive* objective: attributions are computed with respect to the logit difference $f_A(\mathbf{x}) - f_B(\mathbf{x})$ between two confused classes $A$ and $B$. This reveals which tokens push the model toward one genre over the other—a more informative signal than standard single-class IG when genres share substantial vocabulary. We use 30 interpolation steps per sample and aggregate contrastive vocabularies across 30 samples per pair.

# 5 Results

## 5.1 Architecture Comparison

Table 3 summarizes all results. Among the fine-tuned models, **RoBERTa achieves the highest macro F1 of 0.611**, followed by GPT-2 (0.591) and T5 (0.463). However, this ranking must be interpreted with caution due to several confounds:

- RoBERTa's advantage over GPT-2 (2 points F1) is modest and may be partly attributable to its larger trainable classification head (594K vs. 3.8K), providing 150× more task-specific capacity.
- T5's larger gap (−15 points) likely reflects a combination of its smaller base model (`t5-small` at 60M vs. 125M), the additional difficulty of the text-to-text formulation (where decoding errors further degrade performance), and the absence of a trainable classification head.

We emphasize that the ranking RoBERTa > GPT-2 ≫ T5 reflects the *full experimental setup* (architecture + head + base model size), not purely architectural differences. We revisit these confounds in Section 7.4.

Figure 4 visualizes the clustering of all approaches in the 0.46–0.64 F1 range, well above the random baseline (0.200) but well below perfect classification.

## 5.2 Training Dynamics

Figure 5 reveals distinct training behaviors. RoBERTa converges rapidly, reaching peak validation F1 by epoch 2, consistent with bidirectional attention's efficiency at capturing document-level features. GPT-2 requires 7 epochs, consistent with the need for its causal mechanism to gradually build useful representations from left to right. T5 shows the slowest convergence and smallest train–val gap, suggesting it *underfits*—the seq2seq objective adds optimization difficulty that 10 epochs cannot overcome.

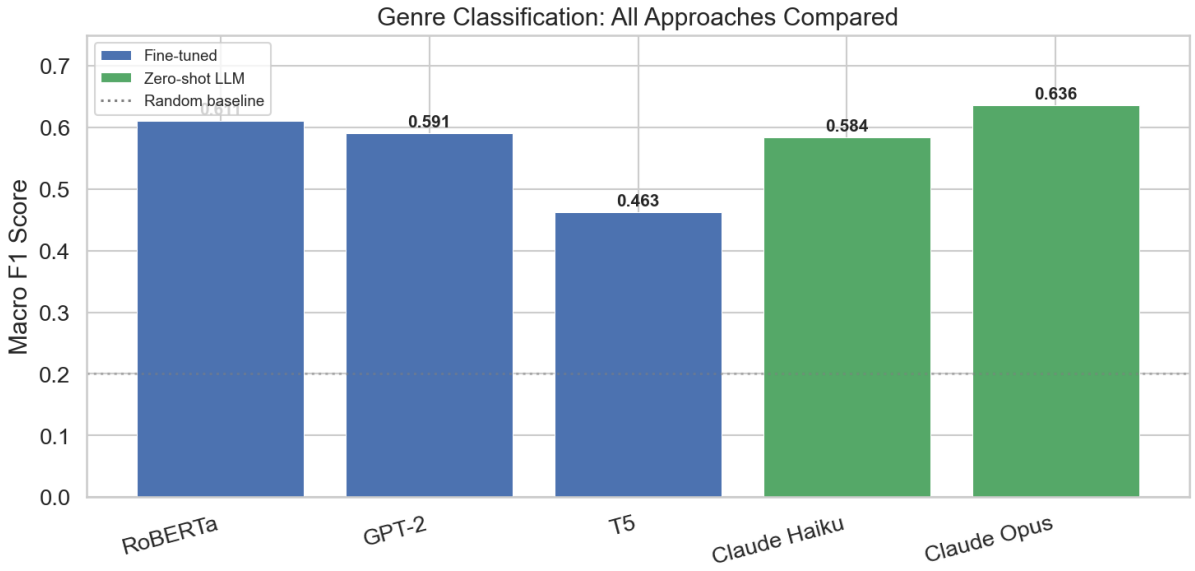



Figure 4: Macro F1 across all approaches. The random baseline is 0.200. All approaches cluster between 0.46–0.64, with Claude Opus setting a soft upper bound at 0.636.

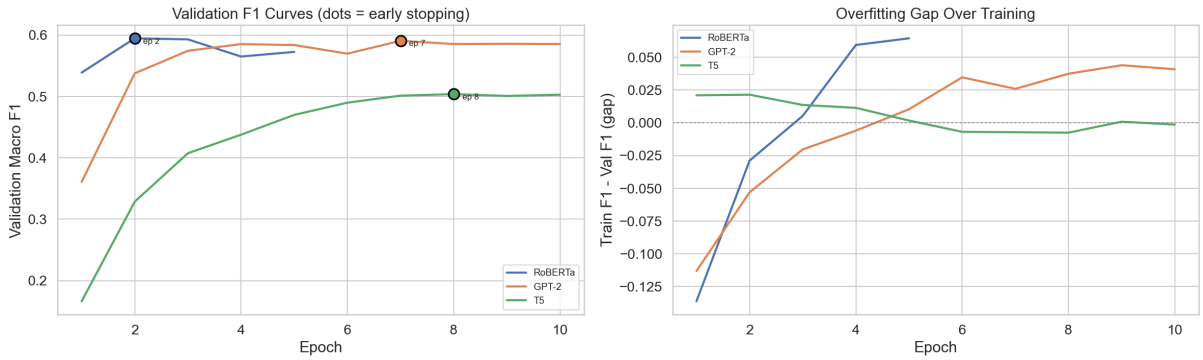



Figure 5: Left: validation F1 curves with early stopping points (dots). RoBERTa peaks at epoch 2; GPT-2 at epoch 7. Right: train–val F1 gap. RoBERTa and GPT-2 overfit progressively; T5 shows minimal gap, suggesting underfitting.

### 5.3 Confusion Analysis

The confusion matrix (Figure 6) reveals systematic patterns. Country is the best-classified genre (82% accuracy), benefiting from distinctive rural themes, while Rap achieves 74% accuracy due to its unique vocabulary. The largest confusion clusters are **Rock→Pop** (70 errors) and **Pop→Country** (50 errors)—precisely the genre pairs with the highest semantic similarity, as we quantify next.

## 6 Understanding the Performance Ceiling

All three fine-tuned models plateau at F1 $\leq 0.61$, and even Claude Opus achieves only 0.636. This section investigates *why* the ceiling exists through four complementary analyses, ordered from the data level (genre similarity) through model behavior (error correlation, embeddings, probing) to token-level explanations (contrastive attribution).

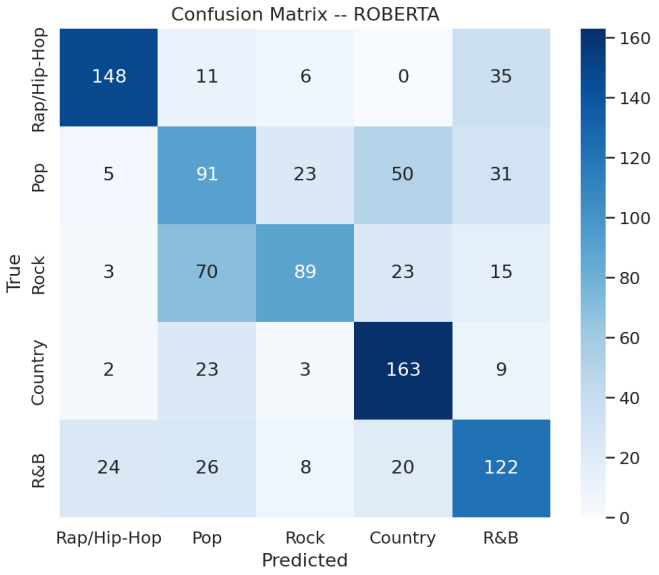



Figure 6: RoBERTa confusion matrix (1,000 test samples, 200 per genre). Country is best-classified (163/200); Pop is most confused, with errors spreading to Rock (23), Country (50), and R&B (31).

### 6.1 Genre Similarity: The Task Is Inherently Hard

Using Sentence-BERT [Reimers and Gurevych, 2019] as an independent embedding model (not our fine-tuned RoBERTa), we find that most genre pairs have centroid cosine similarity above 0.85 (Figure 7). Pop–Rock reaches 0.979 and Pop–Country 0.974—nearly identical in semantic space. Only Rap stands apart (0.84–0.88 with other genres).

At the sample level (Figure 8), within-genre and between-genre similarity distributions overlap heavily (means 0.363 vs. 0.331). This means a given Pop song is often more similar to a random Rock song than to another Pop song—a fundamental constraint on any lyrics-only classifier.

### 6.2 Error–Similarity Correlation: Confusion Mirrors Reality

We connect genre similarity to model behavior by plotting centroid cosine similarity against RoBERTa's confusion rate for each directed genre pair (Figure 9). The Pearson correlation is $r = 0.630$: genres that are more semantically similar are confused more frequently. The three highest confusion rates are:

- Rock→Pop: 35% confusion, similarity 0.979
- Pop→Country: 25% confusion, similarity 0.974
- Rap→R&B: 17.5% confusion, similarity 0.879

Conversely, Rap→Country has 0% confusion and only 0.849 similarity. This establishes that model errors are not random but reflect genuine semantic proximity—the model is *rationally confused*.

### 6.3 Embedding Visualization: Confirming the Overlap

Both UMAP and t-SNE projections (Figure 10) of the learned `[CLS]` representations visually confirm the quantitative findings. Rap forms the most distinct cluster, consistent with its unique vocabulary. Country clusters reasonably well. However, Pop, Rock, and R&B are heavily interleaved, visualizing the genre boundary ambiguity that limits performance.

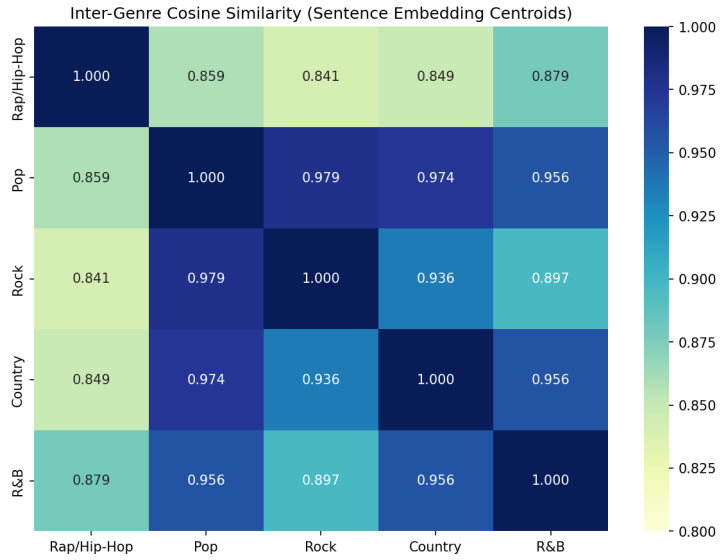



Figure 7: Inter-genre cosine similarity of Sentence-BERT embedding centroids. Pop–Rock (0.979) and Pop–Country (0.974) are nearly indistinguishable, while Rap is the most distinct genre.

### 6.4 Layer Probing: Genre Is a Mid-Level Property

Layer probing (Figure 11) reveals the trajectory of genre information through the network:

- **Layer 0** (static embeddings): F1 = 0.067—barely above random. Word embeddings alone carry negligible genre signal.
- **Layer 1**: F1 jumps to 0.530. A single transformer block integrates enough contextual information for basic genre discrimination.
- **Layers 2–7**: Steady improvement to F1 = 0.613 at layer 7 (the peak).
- **Layers 8–12**: Plateau, with the final layer at F1 = 0.595—slightly *below* the peak.

This trajectory suggests that genre is a *mid-level linguistic property*: deeper than bag-of-words (layer 0 fails) but shallower than the abstract semantic reasoning built in deeper layers. The plateau after layer 7 is a **representational ceiling**: additional depth does not improve genre discrimination because the model has already extracted all linearly accessible signal from the input. This is an *information* ceiling, not a model capacity ceiling.

The per-genre breakdown (Figure 11b) shows that Rap is classifiable from the earliest layers (its vocabulary is highly distinctive), while Pop remains the hardest genre throughout all layers (per-class F1 $\approx$ 0.38), consistent with its weak TF-IDF signature and high overlap with other genres.

### 6.5 Contrastive Integrated Gradients: What Little Signal Exists

The top confused pairs are identified from the confusion matrix (Figure 12). Standard Integrated Gradients answer *which tokens matter for predicting genre A*; contrastive IG answers *which tokens push the model toward A rather than B*. This distinction is critical when genres overlap: a token like "love" has high standard IG for both Pop and Country, but near-zero *contrastive* attribution for the Pop-vs.-Country pair.

For Rock vs. Pop, tokens associated with intense themes (*death*, *blood*, *pain*) push toward Rock, while lighter emotional vocabulary (*love*, *night*, *feel*) pushes toward Pop. Misclassifications

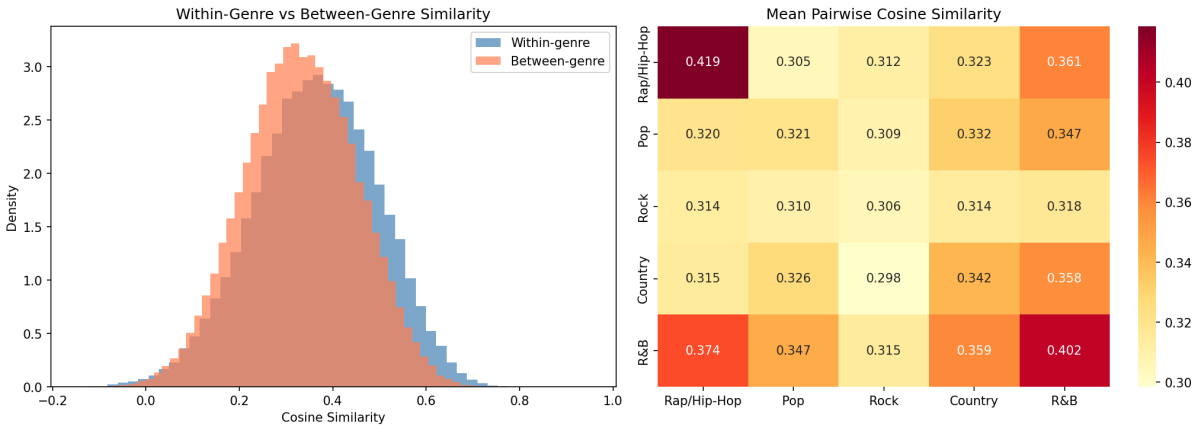



Figure 8: Left: within-genre ($\mu = 0.363$) and between-genre ($\mu = 0.331$) pairwise similarity distributions overlap almost entirely. Right: mean pairwise cosine similarity matrix at the sample level.

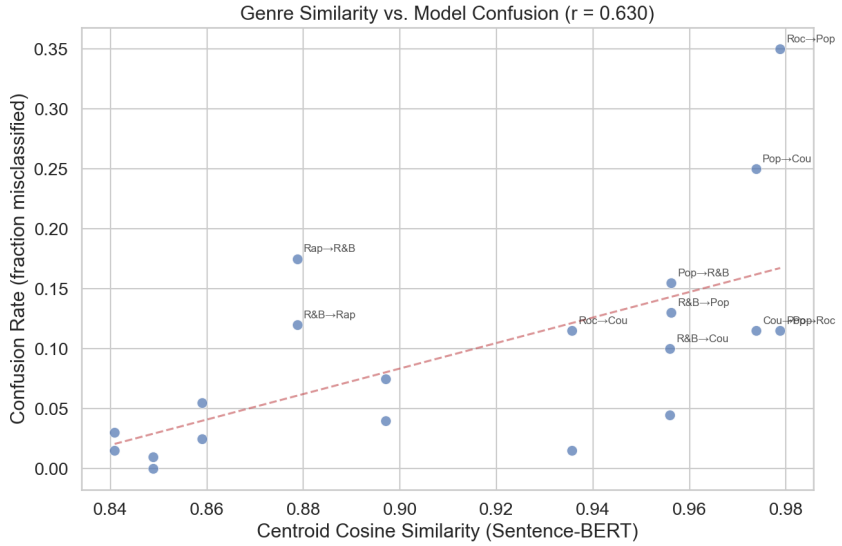



Figure 9: Genre centroid similarity vs. RoBERTa confusion rate for all 20 directed genre pairs (Pearson $r = 0.630$). The most-confused pairs are also the most semantically similar.

occur when a Rock song uses softer language or a Pop song employs darker imagery—genuine lyrical ambiguity.

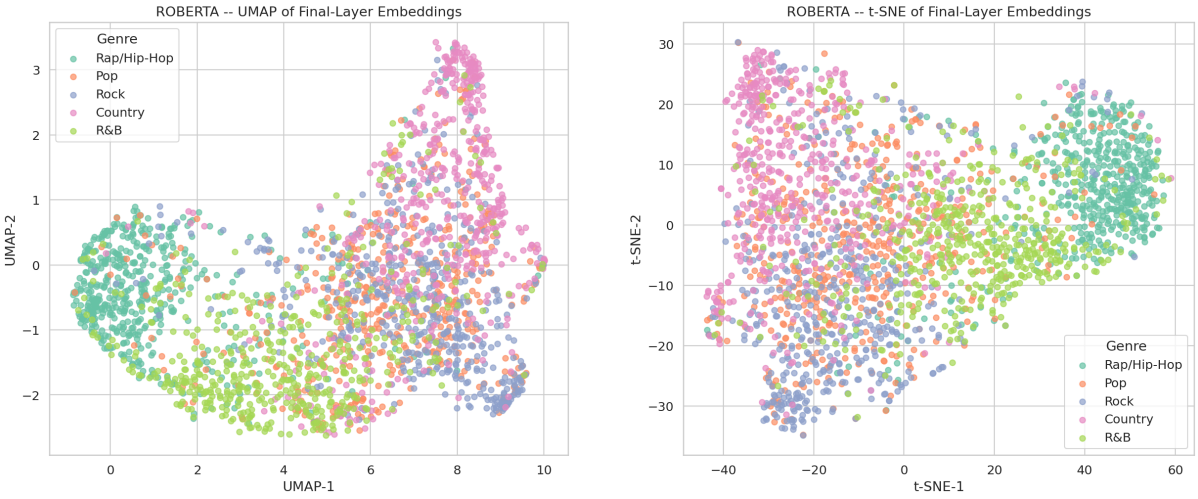
For Country vs. Pop, the model relies on rural/domestic terms (*truck*, *town*, *mama*) to identify Country, but Country-Pop crossover songs often lack these markers. These findings confirm that the model has learned reasonable genre-discriminating features and that errors arise from genuine ambiguity rather than model failure.

# 7 Discussion

## 7.1 The Performance Ceiling

The central finding of this work is that lyrics-only genre classification is bounded at approximately F1 $\approx$ 0.65. Four converging lines of evidence support this conclusion:

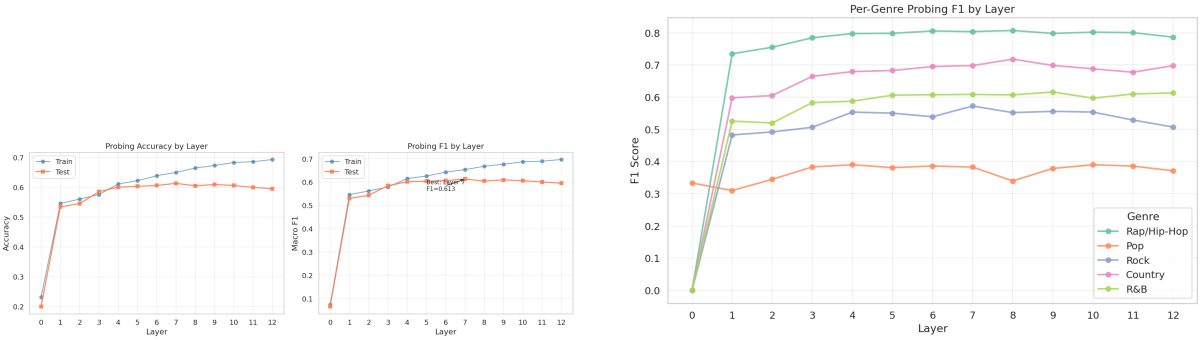
1. **Genre similarity:** Pop–Rock centroid cosine similarity is 0.979 (Sentence-BERT), and within/between-genre distributions overlap almost entirely. These genres are distinguished




(a) UMAP projection.

(b) t-SNE projection.

Figure 10: 2D projections of RoBERTa final-layer `[CLS]` embeddings. Rap and Country form relatively tight clusters; Pop, Rock, and R&B are heavily interleaved.




(a) Overall accuracy and F1 by layer.

(b) Per-genre F1 by layer.

Figure 11: Layer probing results. Genre information is absent at layer 0 (F1 = 0.067), jumps dramatically at layer 1 (0.530), improves through layer 7 (0.613), then plateaus.

primarily by musical properties (instrumentation, tempo, production) that lyrics do not capture.

2. **Layer probing plateau:** Genre information saturates by layer 7, meaning the model has extracted all linearly accessible signal by mid-depth. This is an information ceiling, not a model capacity ceiling.

3. **Error–similarity correlation:** Model confusion mirrors genre proximity ($r = 0.630$), confirming errors arise from data structure, not model failure.

4. **LLM upper bound:** Claude Opus—a frontier LLM with world knowledge of artists and musical context—achieves only F1 = 0.636 in zero-shot mode, only 2.5 points above our fine-tuned RoBERTa.

The marginal gain from world knowledge (0.636 vs. 0.611) suggests that the classification task is fundamentally constrained by the lyrics modality, not by the model's knowledge or capacity.

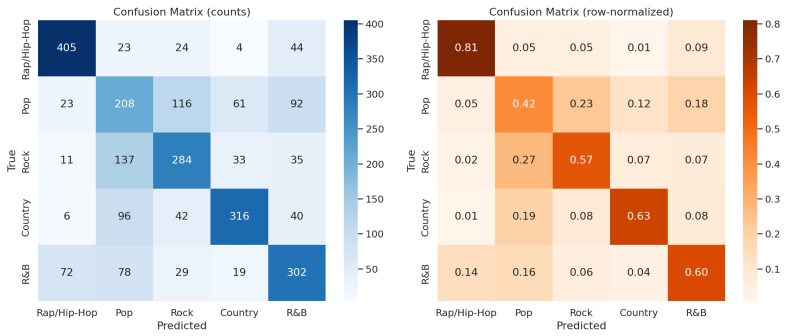



Figure 12: Confusion counts and per-class accuracy for the RoBERTa interpretability model, identifying the top confused pairs.

## 7.2 Scaling Does Not Help

The interpretability RoBERTa, trained on 2.5× more data per genre with augmentation, achieves F1 = 0.608—essentially identical to the comparison RoBERTa (0.611). This is consistent with the ceiling hypothesis: if the input lacks discriminative content, neither more data nor more capacity will help. We note, however, that this Phase 2 experiment changed multiple variables simultaneously (data size, learning rate, augmentation, LoRA targets), so the null improvement cannot be attributed to any single factor.

## 7.3 Architectural Differences vs. Confounds

RoBERTa's advantage over GPT-2 (2 points F1) is consistent with bidirectional attention benefiting document-level classification: the `[CLS]` token can attend to genre cues at arbitrary positions, whereas GPT-2's causal mask forces incremental left-to-right accumulation. However, this comparison is confounded by the 150× larger trainable classification head (594K vs. 3.8K).

T5's larger gap (−15 points) is likely multi-causal: (1) the smaller base model (`t5-small` at 60M vs. 125M) starts with weaker representations; (2) the text-to-text formulation adds unnecessary decoding overhead for a classification task; and (3) T5's language model head is frozen, so all adaptation occurs through LoRA alone. A fairer comparison would use `t5-base` (220M) and equalize head sizes.

## 7.4 Limitations

- **No multi-seed variance reporting.** All results are single-run point estimates. The 2-point F1 gap between RoBERTa and GPT-2 could be within noise. Confidence intervals from multiple random seeds would strengthen or qualify the architectural comparison.
- **Unequal comparison conditions.** The three models differ in base size (60M vs. 125M), total trainable parameters (0.59M vs. 1.18M), and head design. While LoRA adapters are matched (590K), the overall trainable budget is not.
- **No human performance baseline.** Without measuring inter-annotator agreement on lyrics-only genre labeling, we cannot distinguish "the task is inherently hard" from "the genre labels are noisy."
- **Dataset quality.** The Genius dataset contains noise: spam entries, non-lyrical content, and metadata artifacts (see the EDA sample revealing a product description labeled as "Rock").
- **Character truncation.** The 2,000-character limit disproportionately truncates Rap lyrics (median at the cap), potentially removing genre-distinctive content.

- **Genre taxonomy.** Five broad genres is a coarse taxonomy. Finer-grained labels or hierarchical classification could yield different conclusions.
- **LLM sample size.** The Claude evaluation uses 250 samples, introducing higher variance (±6% per-class F1 at 95% confidence) than the full test set.

## 8 Conclusion

We compared three transformer architectures for lyrics-based music genre classification under LoRA fine-tuning. RoBERTa achieved the highest macro F1 (0.611), though confounded by a larger classification head; GPT-2 followed closely (0.591); and T5 underperformed (0.463), partly due to its smaller base model. These rankings reflect the full experimental setup rather than purely architectural differences.

More importantly, we established through four converging analyses—genre similarity, layer probing, error–similarity correlation, and embedding visualization—that lyrics-only classification has an inherent performance ceiling of approximately $F1 \approx 0.65$. This ceiling is data-driven: genres like Pop and Rock are linguistically near-identical (centroid cosine similarity 0.979) and are distinguished primarily by musical properties that lyrics cannot capture. Even Claude Opus, a frontier LLM with world knowledge, achieves only 0.636.

The contrastive Integrated Gradients analysis reveals that when the model does fail, it fails *rationally*: errors occur precisely where lyrics are genuinely ambiguous between genres. This is the core insight: when an information ceiling has been reached, the path forward is not more data or model capacity but richer input modalities. Future work should explore multimodal approaches combining lyrics with audio features to break through this text-only ceiling.

## References


Anthropic. System card: Claude Haiku 4.5. Technical report, Anthropic, October 2025a. URL https://www.anthropic.com/claude-haiku-4-5-system-card.

Anthropic. System card: Claude Opus 4 & Claude Sonnet 4. Technical report, Anthropic, May 2025b. URL https://www-cdn.anthropic.com/4263b940cabb546aa0e3283f35b686f4f3b2ff47.pdf.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loic Barrault, and Marco Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2126–2136, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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*, pages 4171–4186, 2019.

Sebastian Dizon. Genius song lyrics dataset. https://huggingface.co/datasets/sebastiandizon/genius-song-lyrics, 2024.

Michael Fell and Caroline Sporleder. Lyrics-based analysis and classification of music. In *Proceedings of COLING 2014*, pages 620–631, 2014.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Sergio Oramas, Francesco Barbieri, Oriol Nieto, and Xavier Serra. Multimodal deep learning for music genre classification. *Transactions of the International Society for Music Information Retrieval*, 1(1):4–21, 2018.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992, 2019.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 3319–3328, 2017.

Ian Tenney, Dipanjan Das, and Eleni Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, 2019.

Alexandros Tsaptsinos. Lyrics-based music genre classification using a hierarchical attention network. In *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, pages 694–700, 2017.

George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.