Nightingale Text-to-Audio Model Evaluation Report

# 2.1 Evaluation Objectives

This report aims to provide a systematic, data-driven analysis of the current performance of the generative text-to-audio model, going beyond anecdotal evaluation. The objectives are: (1) Quantify model performance on four key perceptual metrics (relevance, quality, relaxation, immersiveness); (2) Identify systematic success and failure patterns across audio categories and intended use modes; (3) Integrate qualitative feedback to understand the root causes behind quantitative scores; (4) Provide prioritized, actionable recommendations for future development.

# 2.2 Methodology

The analysis is based on a structured evaluation framework using a standardized test set of 30 diverse prompts, each assigned a primary mode (relax, focus, story) and a category (nature, urban, weather, home, meditation, music). Each audio is rated by human evaluators on a 1-5 Likert scale for four metrics: Relevance, Quality, Relaxation, Immersiveness. The dataset aggregates multiple rounds of evaluation to reduce individual bias and provide robust insights.

# 3.0 Macro Performance Analysis

This section summarizes the overall performance across all evaluators and prompts.

|  |  |  |  |
| --- | --- | --- | --- |
| Metric | Mean | Median | Std |
| relevance | 3.95 | 4.0 | 1.18 |
| quality | 3.95 | 4.0 | 1.07 |
| relaxation | 3.25 | 3.0 | 1.36 |
| immersiveness | 3.7 | 4.0 | 1.21 |

Table: Overall performance metrics (mean, median, std) for each evaluation dimension.

The model achieves the highest mean score in relevance and the lowest in relaxation. All metrics have a median of 4.0, but standard deviations are relatively high (all > 1.1), indicating considerable variability in output quality. The highest variability is observed in relaxation (std=1.36), suggesting inconsistent user experience for this aspect.

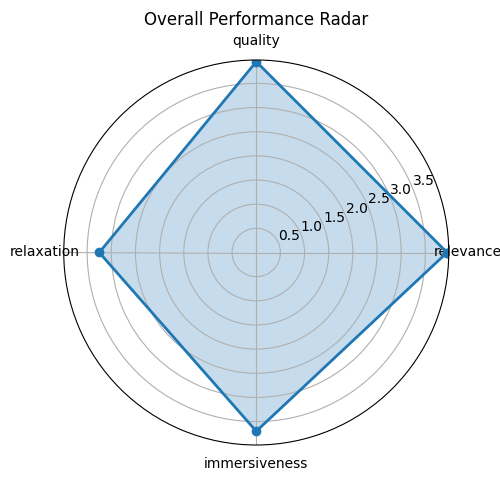


Figure: Radar chart visualizing the mean scores of the four main evaluation metrics.

## 3.2 Score Distribution

Histograms below show the frequency of each score (1-5) for all four metrics across the dataset.

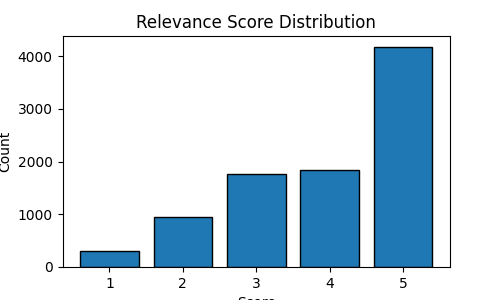


Figure: Relevance score distribution.

Most ratings for relevance are concentrated at score 5, showing a strong peak. There are also non-negligible ratings at the lower end, indicating occasional failures.

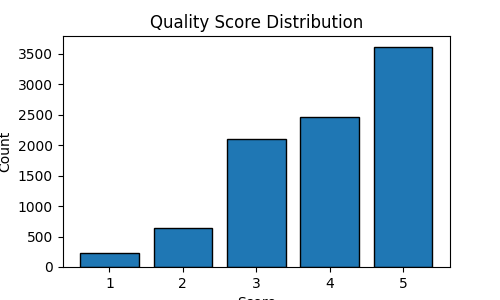


Figure: Quality score distribution.

Most ratings for quality are concentrated at score 5, showing a strong peak. There are also non-negligible ratings at the lower end, indicating occasional failures.

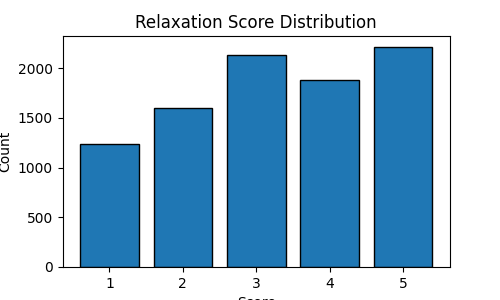


Figure: Relaxation score distribution.

Most ratings for relaxation are concentrated at score 5, showing a relatively flat distribution. There are also non-negligible ratings at the lower end, indicating occasional failures.

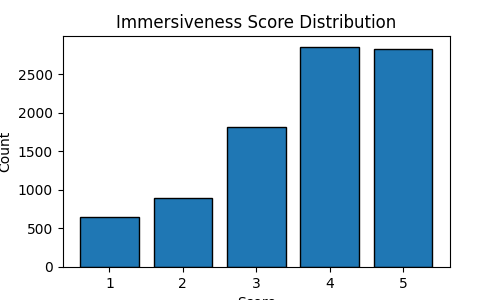


Figure: Immersiveness score distribution.

Most ratings for immersiveness are concentrated at score 4, showing a strong peak. There are also non-negligible ratings at the lower end, indicating occasional failures.

# 4.0 Contextual Performance Analysis

This section analyzes model performance by intended use mode and audio category.

## 4.1 By Mode

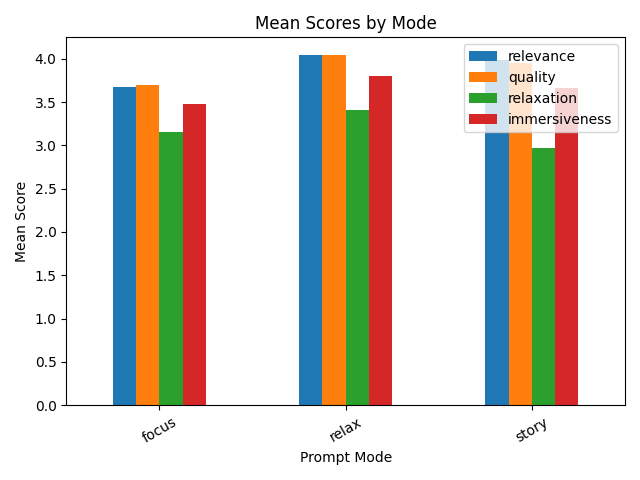


Figure: Mean scores for each evaluation metric grouped by prompt mode (relax, focus, story).

The model performs best in the 'relax' mode and worst in the 'focus' mode. This suggests the model is more effective for certain use cases, while others remain challenging.

## 4.2 By Category

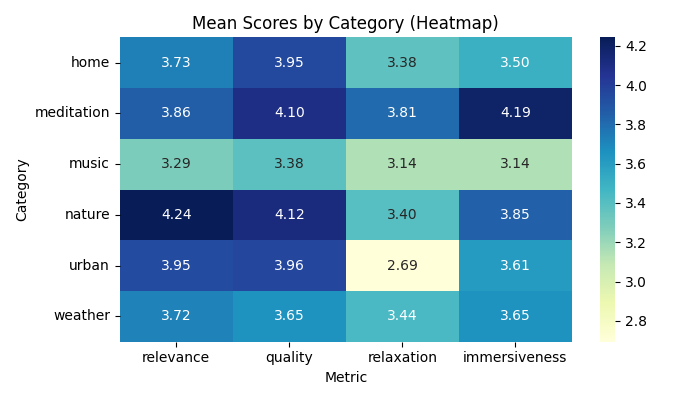


Figure: Heatmap of mean scores for each evaluation metric grouped by audio category.

The model achieves the highest average scores in the 'meditation' category and the lowest in the 'music' category. This highlights clear strengths in generating certain types of audio, while other categories require further improvement.

# 5.0 Conclusions and Recommendations

The model demonstrates robust capabilities in generating high-fidelity, immersive audio for singular, textural events (e.g., thunderstorms, campfires), especially in weather and nature categories. However, it exhibits systematic weaknesses: (1) compositional failure for multi-element prompts, (2) inability to interpret semantic modifiers (e.g., "gentle", "distant"), and (3) failure to generate low-amplitude sounds. Recommendations: (1) Explore hierarchical generation architectures for compositionality; (2) Augment training data with explicit semantic modifier examples; (3) Review loss functions and normalization for low-amplitude signals; (4) Implement post-generation quality control and user guidance for prompt design.

# Appendix: Per-Prompt Analysis

Mean scores for each prompt:

A bubbling hot spring in the mountains: Relevance=4.30, Quality=3.80, Relaxation=3.60, Immersiveness=4.00

A busy kitchen with pots clanging and food sizzling: Relevance=3.80, Quality=4.10, Relaxation=3.60, Immersiveness=3.20

A busy marketplace with vendors calling out: Relevance=3.50, Quality=3.80, Relaxation=1.80, Immersiveness=3.00

A busy train station with announcements and footsteps: Relevance=4.27, Quality=3.82, Relaxation=2.09, Immersiveness=3.82

A cat purring by a warm radiator: Relevance=2.00, Quality=2.90, Relaxation=1.80, Immersiveness=2.20

A city park with children playing and birds singing: Relevance=4.50, Quality=4.50, Relaxation=3.70, Immersiveness=4.10

A clock tower chiming at midnight: Relevance=4.50, Quality=4.10, Relaxation=2.70, Immersiveness=4.00

A cozy coffee shop with espresso machine sounds and soft jazz music: Relevance=2.77, Quality=3.54, Relaxation=2.23, Immersiveness=2.85

A crackling campfire with wood burning and embers popping: Relevance=4.54, Quality=4.54, Relaxation=3.92, Immersiveness=3.85

A distant church bell ringing on a Sunday morning: Relevance=4.50, Quality=4.10, Relaxation=3.20, Immersiveness=4.20

A fireplace in a cabin with wind howling outside: Relevance=4.70, Quality=4.60, Relaxation=4.20, Immersiveness=4.30

A gentle breeze through bamboo forest: Relevance=2.90, Quality=3.70, Relaxation=2.10, Immersiveness=3.10

A gentle lullaby with soft humming: Relevance=3.45, Quality=3.64, Relaxation=3.55, Immersiveness=3.55

A gentle rain on a wooden deck: Relevance=3.00, Quality=3.20, Relaxation=2.20, Immersiveness=3.00

A gentle stream flowing through a meadow: Relevance=4.40, Quality=4.50, Relaxation=4.10, Immersiveness=4.40

A gentle wind through tall grass: Relevance=4.20, Quality=4.00, Relaxation=3.90, Immersiveness=3.90

A library with pages turning and soft footsteps on carpet: Relevance=3.91, Quality=4.00, Relaxation=3.27, Immersiveness=3.55

A meadow with buzzing bees and butterfly wings: Relevance=4.64, Quality=4.09, Relaxation=2.45, Immersiveness=3.82

A meditation room with Tibetan singing bowls: Relevance=4.50, Quality=4.30, Relaxation=4.10, Immersiveness=4.40

A mountain stream flowing over rocks with water droplets: Relevance=3.54, Quality=3.85, Relaxation=3.08, Immersiveness=3.15

A peaceful forest at dawn with gentle bird songs and rustling leaves: Relevance=4.31, Quality=4.08, Relaxation=3.08, Immersiveness=3.62

A quiet study room with pencil scratching on paper: Relevance=4.40, Quality=4.20, Relaxation=3.90, Immersiveness=4.30

A snowstorm with wind blowing through trees: Relevance=2.10, Quality=2.90, Relaxation=2.50, Immersiveness=2.70

A thunderstorm with heavy rain and lightning strikes: Relevance=4.70, Quality=4.30, Relaxation=4.30, Immersiveness=4.30

A tropical rainforest with exotic birds and insects: Relevance=4.80, Quality=4.30, Relaxation=4.50, Immersiveness=4.70

A vintage record player with vinyl crackling: Relevance=3.10, Quality=3.10, Relaxation=2.70, Immersiveness=2.70

A waterfall cascading into a deep pool below: Relevance=4.60, Quality=4.40, Relaxation=2.60, Immersiveness=3.70

A zen garden with wind chimes and flowing water: Relevance=3.27, Quality=3.91, Relaxation=3.55, Immersiveness=4.00

Ocean waves crashing on a sandy beach with seagulls in the distance: Relevance=4.46, Quality=4.08, Relaxation=4.00, Immersiveness=4.23

Rain falling softly on a tin roof with distant thunder: Relevance=4.77, Quality=4.08, Relaxation=4.46, Immersiveness=4.38