3. Evaluation of Gen Music Systems

Generative Algorithms for Sound and Music

End-to-end pipeline

1. Problem definition
2. Data strategy
3. Representation & pre-processing
4. Model choice
5. Training / implementation
6. Evaluation
7. Deployment

End-to-end pipeline

1. Problem definition
2. Data strategy
3. Representation & pre-processing
4. Model choice
5. Training / implementation
6. Evaluation
7. Deployment

Evaluation is the most important part of gen music,

but the most underdeveloped

Overview

1. The evaluation problem
2. Pragmatic evaluation
3. Families of evaluation metrics
4. Evaluation across development stages

1. The evaluation problem

The evaluation problem

- Many papers present systems with:
  - No evaluation
  - Minimal qualitative examples
  - Cherry-picked samples
- Negligence + structural problem

Why is it so hard to evaluate a gen music system?

The evaluation problem

The evaluation problem

- Music has no objective goal
  - Chess → win / lose
  - Classification → accuracy
  - Music → what is success?

The evaluation problem

- Music has no objective goal
  - Chess → win / lose
  - Classification → accuracy
  - Music → what is success?
- Aesthetic judgment is deeply subjective
  - Cultural background
  - Genre familiarity
  - Personal taste

The evaluation problem

- Music has no objective goal
  - Chess → win / lose
  - Classification → accuracy
  - Music → what is success?
- Aesthetic judgment is deeply subjective
  - Cultural background
  - Genre familiarity
  - Personal taste
- Massive variance in perception
  - What sounds boring to one listener is profound to another

What should we evaluate?

- Music quality?
- Plausibility of the output?
- Adherence to a style?
- Degree of creativity of the output?
- The level of creativity of the system?
- …

We only improve what we can measure

2. Pragmatic evaluation

Evaluation for AI music companies

- Affects iteration speed
- Product decisions
- Revenue and retention
- Life or death

Pragmatic approach to evaluation

1. Leave aside philosophical / musicological debates

Pragmatic approach to evaluation

1. Leave aside philosophical / musicological debates
2. Wear an engineer hat

Pragmatic approach to evaluation

1. Leave aside philosophical / musicological debates
2. Wear an engineer hat
3. Accept we cannot measure *music quality* directly

Pragmatic approach to evaluation

1. Leave aside philosophical / musicological debates
2. Wear an engineer hat
3. Accept we cannot measure *music quality* directly
4. Rely on *proxy metrics*

Proxy metric mindset

- A proxy metric:
  - Doesn’t measure the true goal
  - Measures something correlated with it

Proxy metric mindset

- A proxy metric:
  - Doesn’t measure the true goal
  - Measures something correlated with it
- Each metric captures one projection of a complex phenomenon

Proxy metric mindset

- A proxy metric:
  - Doesn’t measure the true goal
  - Measures something correlated with it
- Each metric captures one projection of a complex phenomenon
- No single metric is sufficient

Proxy metric mindset

- A proxy metric:
  - Doesn’t measure the true goal
  - Measures something correlated with it
- Each metric captures one projection of a complex phenomenon
- No single metric is sufficient
- Goal is to:
  - Select a small set of meaningful proxies
  - Understand what each proxy can and cannot tell you

Core evaluation principle

Evaluation

=

Triangulation

3. Families of evaluation metrics

Four families of evaluation metrics

- Objective (symbolic, audio)
- Subjective / Human perceptual evaluation
- Expert-based
- Market-based

Objective metrics

- Compare generated output against dataset
- Hope for similarity
- Metric = deterministic, objective, computational musicology

Symbolic objective metrics

- Pitch class distribution
- Interval distributions
- Key consistency over time
- Chord distribution
- Note density
- Inter-onset interval (IOI) distributions
- Repetition rates
- Motif recurrence
- …

Symbolic objective metrics

- Pitch class distribution
- Interval distributions
- Key consistency over time
- Chord distribution
- Note density
- Inter-onset interval (IOI) distributions
- Repetition rates
- Motif recurrence
- …

Pitch

Harmony

Rhythm

Structure

Audio objective metrics

- Fréchet Audio Distance (FAD)
  - Measures distance between embeddings of real vs generated audio
  - Inspired by FID in images
  - Lower = closer to real data distribution

Audio objective metrics

- Fréchet Audio Distance (FAD)
  - Measures distance between embeddings of real vs generated audio
  - Inspired by FID in images
  - Lower = closer to real data distribution
- Embedding-space similarity
  - Compare embeddings from pretrained models (e.g. MERT-like models)
  - Measures statistical similarity, not musical quality

R&D team: “Music has improved after our last training run.”

V: “How do you know that?”

R&D team: “FAD is 20% lower.”

R&D tea d after our last traini

V: “How do

R&D team: ”

Objective metrics strengths

Objective metrics strengths

- Fast
- Cheap
- Repeatable
- Ideal for:
  - Early-stage iteration
  - Regression testing
  - Detecting obvious failures

Objective metrics limitations

Objective metrics limitations

- Do **not** tell you if music is good
- Easy to game
- Reward statistical mimicry over creativity
- Correlate poorly with human enjoyment

Subjective evaluation

- Asking humans to listen and judge generated music
- Closest approximation to real listening experience
- Amazon Mechanical Turk

Subjective evaluation: Common methods

- Mean Opinion Score (MOS) / Likert scales
  - “Rate quality from 1 to 5”
  - Simple and common

Subjective evaluation: Common methods

- Mean Opinion Score (MOS) / Likert scales
  - “Rate quality from 1 to 5”
  - Simple and common
- A/B preference tests
  - “Which sample do you prefer?”
  - Comparative judgment across models
  - Reduces individual rating bias

Subjective evaluation: Common methods

- Mean Opinion Score (MOS) / Likert scales
  - “Rate quality from 1 to 5”
  - Simple and common
- A/B preference tests
  - “Which sample do you prefer?”
  - Comparative judgment across models
  - Reduces individual rating bias
- Task-based evaluation
  - “Would you use this in a game / video / playlist?”
  - Contextualizes judgment

Subjective evaluation strengths

Subjective evaluation strengths

- Closest to how music is actually consumed
- Captures holistic perception
- Sensitive to musical coherence

Subjective evaluation limitations

Subjective evaluation limitations

- Noisy
- Expensive
- Sensitive to:
  - Instructions
  - Framing
  - Listener fatigue
  - Order effects
- Hard to reproduce exactly


NOT A BUG, A FEATURE
HUMANS DISAGREE BECAUSE MUSIC PERCEPTION IS DIVERSE
imgflip.com

Expert evaluation

- Evaluation by trained musicians, composers, producers
- Focuses on *why* something works or fails

Expert evaluation

- Evaluation by trained musicians, composers, producers
- Focuses on *why* something works or fails
- Evaluation criteria:
  - Structural coherence
  - Stylistic authenticity
  - Voice leading and harmony
  - Musical interest vs cliché
  - Long-term form
  - Failure mode identification

Expert evaluation strengths

Expert evaluation strengths

- High signal-to-noise ratio
- Deep qualitative insights
- Can diagnose problems models cannot reveal

Expert evaluation limitations

Expert evaluation limitations

- Expensive
- Low scalability
- Hard to standardize
- Experts may disagree strongly
- Biased toward specific aesthetics or traditions

Market-based evaluation

- Measuring what users actually do in the real world
- Behavior not opinion

Market-based evaluation

- Measuring what users actually do in the real world
- Behavior not opinion
- Typical metrics
  - Skip rate
  - Retention
  - Repeated usage
  - Likes / saves
  - Conversion (free → paid)
  - Session duration

Market-based evaluation strengths

Market-based evaluation strengths

- Directly tied to impact
- Scales naturally
- Hard to fake

Market-based evaluation limitations

Market-based evaluation limitations

- Optimizes usefulness, not aesthetics
- Depends heavily on UX, branding, context
- Can reward mediocrity over originality


MARKET SUCCESS
≠ MUSICAL EXCELLENCE
宣王
MARKET FAILURE USUALLY MEANS NO REAL-WORLD VALUE
imgflip.com

Activity: Design an evaluation strategy

- Work in groups of 3 or 4 (10 minutes)
- Address 2 use cases:
  - MelodyStudio by WaveAI
  - RAVE (Real-time Audio Variational autoEncoder)
- Decide
  - Primary goal of eval
  - Evaluation methods / metrics + why?
  - One method you would not trust

4. Evaluation across development stages

Evaluation across development stages

- Evaluation types are not interchangeable
- Each fits a different phase

Early-stage development

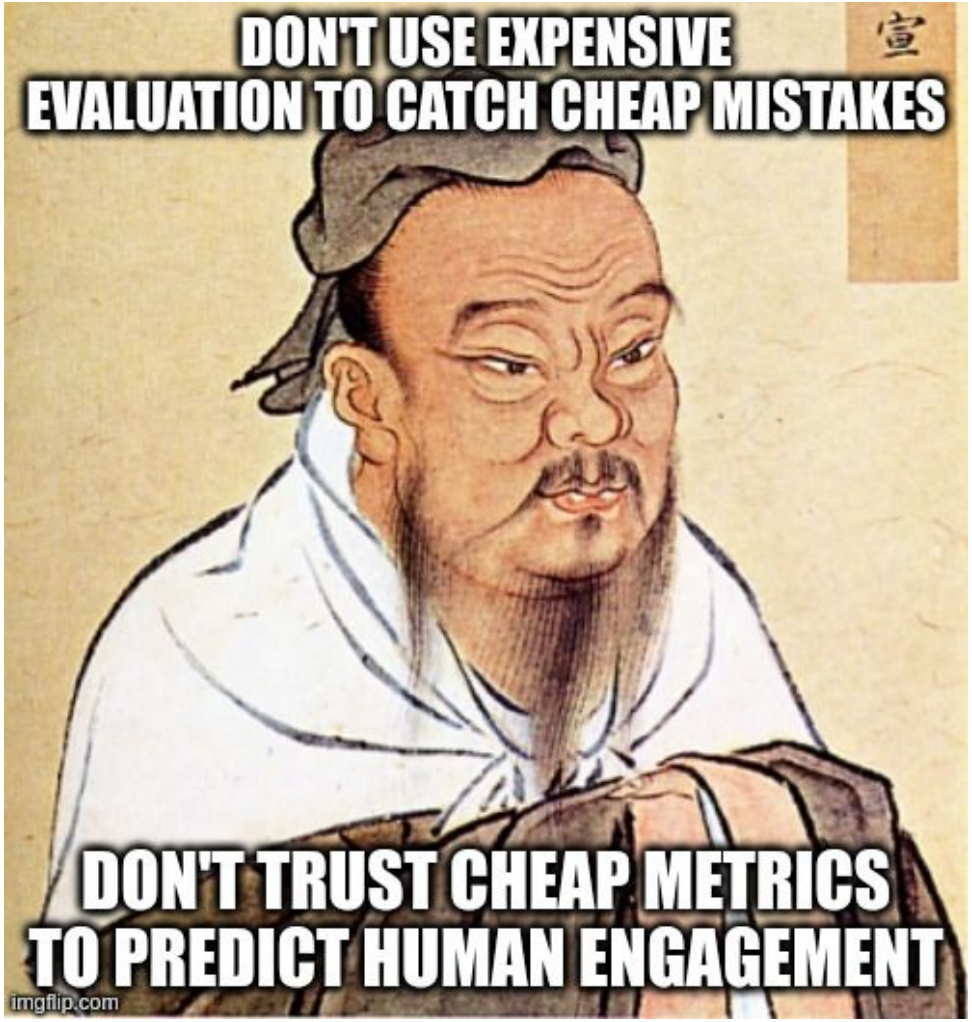
- Objective metrics
- Regression tests
- Sanity checks

Mid-stage validation

- Human perceptual evaluation
- Expert feedback
- Model comparison

Product phase

- Market metrics
- Retention and engagement
- Iterative optimization


DON'T USE EXPENSIVE
EVALUATION TO CATCH CHEAP MISTAKES
DON'T TRUST CHEAP METRICS
TO PREDICT HUMAN ENGAGEMENT
imgflip.com

Which metric is best?

~~Which metric is best?~~

What am I trying to optimize for?

Evaluation metrics are context dependent

- Creative tool → expert + human evaluation
- Background music product → market metrics
- Internal iteration → objective metrics

Takeaways

Takeaways

- Metrics are proxies

Takeaways

- Metrics are proxies
- You can draw from many metric families:
  - Objective
  - Subjective
  - Expert
  - Market-based

Takeaways

- Metrics are proxies
- You can draw from many metric families:
  - Objective
  - Subjective
  - Expert
  - Market-based
- Always align evaluation with:
  - System goals
  - Development stage
  - Real-world use case