SoundSafe.AI Watermark Model Training and Compute Plan

1. MVP Baseline Overview

The MVP watermarking model had the following characteristics:

- **Dataset**: ~5,000 hours of audio (16 kHz).
- **Training Steps**: ~70,000 steps.
- **Limitations**:
  - Audio resolution limited to 16 kHz.
  - Weak robustness against manipulations.
  - Inefficiency with high-payload metadata.
  - No real-time processing capability.

The MVP trained for **core encoding**, **attack robustness**, and **perceptual quality fine-tuning** across three stages.

2. Goals for Commercial-Grade Model

The upgraded model is designed to achieve the following:

1. **Support Higher Audio Quality**: Move to 44.1 kHz or higher resolution.
2. **Enable Real-Time Processing**: Process audio chunks while handling long-range dependencies.
3. **Enhance Robustness**: Resist multiple, layered attacks.
4. **Support Complex Metadata Payloads**:
   - Encode 18 metadata fields (DDEX/CWR standards) with variable lengths.
   - Handle dynamic, hierarchical metadata structures.
   - Optimize for payloads requiring ~27,792 bits (~3,474 bytes).
5. **Scalability**: Handle large datasets of 100,000+ hours (~100TB).

3. Estimated Training Requirements

Training Steps

The commercial-grade model requires significantly more steps due to increased complexity:

- **Audio Complexity (44.1 kHz++):** ~3x increase in data points per second.
- **Real-Time/Long-Length Training:** ~2x increase to address long-range dependencies.
- **Enhanced Robustness:** ~2x increase for handling diverse manipulations.
- **Increased Metadata Payload:** ~1.5x increase for hierarchical encoding.

Baseline Calculation:

```
New Steps = 70,000 (MVP Steps) * 3 * 2 * 2 * 1.5 ≈ 1.26M steps
```

Additional Payload Optimizations

- **Dynamic Bit Encoding**: +10,000 steps.
- **Variable-Length Payload Training**: +20,000 steps.
- **Metadata Representation Learning**: +30,000 steps.

Revised Total Steps:

```
1.26M + 60,000 ≈ 1.32M steps
```

Dataset

- **Target Size**: ~100,000 hours (~100TB) of audio.

- **Augmentations**:
    - Simulated manipulations (compression, EQ, time-stretching, etc.).
    - Variable-length metadata subsets.
    - Edge cases (maximum character fields, erroneous metadata).

4. Training Phases and Compute Breakdown

Training Phases

Phase	Steps	Goals
Core Metadata Training	500,000 steps	Encode high-priority fields (e.g., ISRC) while ensuring minimal distortion.
Variable-Length Metadata	300,000 steps	Support dynamic and hierarchical metadata fields.
Hierarchical Metadata	250,000 steps	Train for chunked, sequential encoding and variable bit allocation.
Loss Function Fine-Tuning	120,000 steps	Optimize loss weighting for high-priority fields to reduce error rates.
Real-Time Inference Training	150,000 steps	Enable accurate, robust real-time processing of 1–5 second audio chunks.

Total Steps: **1.32M steps**

Compute Resources

VM Selection

Task	VM Type	Purpose	Cost (Spot Pricing)
Base Training	NDm_A100 v4	Parallel training with large batches	~$10.04/hour
Fine-Tuning and Inference	NCas_T4 v3	Cost-efficient for smaller datasets	~$1.5/hour
Dataset Preparation/Validation	D8_v5	Preprocessing and metadata validation	~$0.7/hour

Cost Breakdown

Phase	Steps	VM Type	Estimated Cost
Core Metadata Training	500,000 steps	NDm_A100 v4	$35,000–$40,000

Variable-Length Metadata	300,000 steps	NDm_A100 v4	$20,000–$25,000
Hierarchical Metadata	250,000 steps	NDm_A100 v4	$15,000–$20,000
Loss Function Fine-Tuning	120,000 steps	NCas_T4 v3	$2,500–$3,000
Real-Time Inference Training	150,000 steps	NCas_T4 v3	$3,500–$4,000
Validation	~1,000 hours	D8_v5	~$700–$1,000

Total Compute Cost: **$76,700–$93,000**

5. Optimization Strategies

Compute Optimizations

1. **Dynamic Scaling**: Begin with smaller datasets and scale as the model converges.
2. **Mixed Precision Training**: Use FP16 to reduce GPU memory usage (~50% reduction).
3. **Distributed Training**: Utilize Azure Machine Learning for efficient scaling.
4. **Early Stopping**: Terminate training early to save time and costs.

Model Training Optimizations

1. **Curriculum Learning**: Progressively train from fixed-length to hierarchical payloads.
2. **Augmentation**: Include manipulations like compression and clipping for robustness.
3. **Loss Function Tuning**: Prioritize high-value fields like ISRC.
4. **Checkpoints**: Save intermediate checkpoints to prevent retraining.

6. Summary of Key Metrics

- **Training Steps**: ~1.32M steps.
- **Dataset Size**: ~100,000 hours (~100TB).
- **Compute Cost**: $76,700–$93,000.
- **Target Output**:
  - Support for **44.1 kHz++ audio resolution**.
  - Real-time processing of **1–5 second chunks**.
  - Metadata recovery for payloads up to **3,474 bytes**.