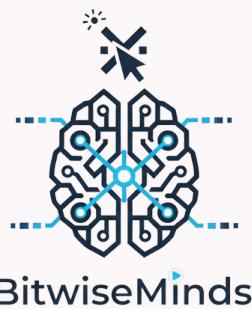


# শব্দতরী: Where Dialects Flow into Bangla

Nov 18, 2025



## TEAM BITWISEMINDS

Presented by Samiul Basir Bhuiyan, Md. Sazzad Hossain Adib, Mohammed Aman Bhuiyan

# 01

## TRAINING SUMMARY

---

### Dataset Preparation

<https://huggingface.co/datasets/bitwisemind/hackathon>

- **Total Samples:** 3,350 audio files with transcriptions
- **Train/Test Split:** 90/10 ratio (3,015 training / 335 testing samples)
- **Audio Format:** WAV files resampled to 16kHz
- **Preprocessing:** Audio converted to log-Mel spectrogram features, transcriptions tokenized to label IDs

### Model Configuration

- **Base Model:** We finetuned [bengaliAI/tugstugi\\_bengaliai-regional-asr\\_whisper-medium](https://huggingface.co/bengaliAI/tugstugi_bengaliai-regional-asr_whisper-medium) model which originally [openai/whisper-medium](https://huggingface.co/openai/whisper-medium) model, finetuned in [Ben10](#) Dataset.
- **Architecture:** Whisper Medium (Sequence-to-Sequence Transformer)
- **Feature Extraction:** WhisperFeatureExtractor with 16kHz sampling rate
- **Tokenizer:** WhisperTokenizer configured for Bengali transcription

### Training Configuration & Evaluation

- **Batch Size:** 4 per device
- **Gradient Accumulation Steps:** 4 (Effective batch size: 16)
- **Learning Rate:** 1e-5 with 200 warmup steps
- **Total Training Steps:** 4,000 steps (~8.5 hours)
- **Optimization:** AdamW optimizer with gradient clipping (max\_norm=1.0)
- **Precision:** FP16 mixed precision training (O1 level)
- **Memory Optimization:** Gradient checkpointing enabled
- **GPU:** Single Tesla T4 (15GB VRAM)
- **Evaluation Metric:** WER (Word Error Rate) - lower is better
- **Evaluation Frequency:** Every 1,000 steps

# 02

## INFERENCE SUMMARY

---

### Model Loading

- **Total Samples:** 3,350 audio files with transcriptions
- **Train/Test Split:** 90/10 ratio (3,015 training / 335 testing samples)
- **Audio Format:** WAV files resampled to 16kHz
- **Preprocessing:** Audio converted to log-Mel spectrogram features, transcriptions tokenized to label IDs

### Model Configuration

- **Model:** [bitwisemind/bitwisemind-whisper-medium-bangla](#)

### Audio Preprocessing

- **Input Format:** WAV files (16kHz sample rate)
- **Loading:** librosa-based audio loading with automatic resampling
- **Feature Extraction:** Whisper processor converts audio to input features

### Inference Configuration

- Max length: 225 tokens
- Beam search: 5 beams
- Special tokens: Automatically skipped in decoding

### Pipeline Flow

- Scan and validate audio files
- Load pre-trained model and processor
- Process each audio file sequentially with progress tracking
- Generate transcriptions using beam search
- Store results with error handling
- Export to CSV with comprehensive statistics