Project Koenigstiger

Whisper German Fine-Tune

Model Card | Version 1.0 | Date: January 22, 2026

openml.io

"Loquor, ergo sum"

Executive Summary

The **Koenigstiger German ASR** model is a fine-tuned version of OpenAI's `whisper-small`. It is specifically optimized for high-accuracy transcription of German speech in noisy, spontaneous environments (movies, interviews, unstructured dialogue), bridging the gap between "read speech" datasets and real-world audio.

Model Details

Developer: OpenML (Project Koenigstiger)
Model Architecture: Transformer (Encoder-Decoder)
Base Model: `openai/whisper-small` (244M params)
Input: Raw Audio (16kHz Mono)
Output: German Text Transcripts
License: Apache 2.0 (Derived)

Release Date: January 2026
Frameworks: PyTorch, Hugging Face Transformers
Storage Format: Hugging Face `.safetensors`
Inference Engine: CTranslate2 / Faster-Whisper
Hardware: NVIDIA Tesla T4 (AWS g4dn.xlarge)

Training Data

The model was trained using a hybrid "Base + Specialist" strategy to mitigate domain drift.

Dataset Source	Description	Role
Mozilla Common Voice	30GB+ of crowdsourced "Read Speech". Used to establish grammatical and phonemic fluency in standard German	Foundation
Local Video Shards	Proprietary collection of movie clips and interviews. Contains overlapping dialogue, background noise, and colloquial slang.	Specialization

Training Procedure

The training utilized a shard-based continual learning pipeline to handle dataset sizes exceeding local disk capacity.

- **Data Sharding:** The massive source dataset (30GB+) was pre-processed into self-contained 1GB shards (.tar archives containing paired audio and transcripts) to facilitate streaming ingestion without requiring the entire dataset to be extracted at once.
- **Pipeline:** Sequential Shard Loading (Load $\rightarrow$ Train $\rightarrow$ Save $\rightarrow$ Purge).
- **Precision:** Mixed Precision (FP16) to optimize memory on 16GB VRAM.
- **Optimization Strategy:**
  - **Shards 1-2:** Learning Rate `1e-5`. Aggressive adaptation to German vocabulary.
  - **Shards 3+:** Learning Rate `1e-6`. Conservative updates to prevent catastrophic forgetting.
- **Batch Size:** Effective batch size of 16 (4 per device $\times$ 4 gradient accumulation steps).

Evaluation Results

We evaluated the model against a "Gold Standard" test set containing difficult movie dialogue (e.g., *"Scharfführer..."*).

Model Version	Observation / Output Quality	Relative Accuracy
Stock Whisper-Base	Struggled with proper nouns and military ranks.	Baseline (81%)
Koenigstiger v1 (Shard 2)	**Correctly identified "Scharfführer" and "Malheur".** Optimal balance of vocabulary and acoustic robustness.	**Peak (91%)**
Koenigstiger v1 (Shard 3)	Accuracy plateaued despite additional training. Training was halted to prevent overfitting to "Read Speech" style.	Plateau (91%)

Limitations & Ethical Considerations

1. **Language Bias:** The model is heavily biased towards German. If fed English audio, it may attempt to transcribe it phonetically as German.
2. **Domain Drift:** While improved for movies, the model may still struggle with extreme background music or overlapping shouting matches compared to studio recording.
3. **Usage:** Intended for educational purposes (language learning) and accessibility. Not validated for medical or legal transcription.

Contact: jack20220723@gmail.com | **Repository:** N/A (private model)