

TRANSCRIPTION-FIRST PIANO DIFFICULTY ESTIMATION WITH A FINE-TUNED SYMBOLIC FOUNDATION MODEL

Jai Dhiman

Independent Researcher

jaidhiman2000@gmail.com

ABSTRACT

We describe CrescendAI, our submission to the MIREX 2026 Music Performance Difficulty Prediction task. The system is transcription-first: audio is converted to MIDI with Transkum, an automatic piano transcription model, and the resulting symbolic sequence is encoded by MoonBeam-839M, a MIDI-native foundation model, fine-tuned with LoRA adapters; token embeddings are mean-pooled over the whole piece, and a ridge head emits one real-valued difficulty score. The submitted system averages the predictions of five adapters that differ only by training seed. Training used 7,665 transcriptions of the public Piano Syllabus corpus; no other data was used, and the evaluation repertoire is disjoint from that corpus by the task’s own construction. On composer-disjoint cross-validation the recipe scores Kendall’s τ_c 0.8430 for a single adapter and 0.8461 for the five-adapter ensemble, against 0.8068 for a supervision-matched hand-feature baseline on identical folds, and within 0.024 of an estimated label-noise ceiling of τ_c 0.8668. The submitted model applies that same recipe to all 7,665 pieces with nothing held out, so we report no score for it: any evaluation would be train-on-test.

1. SYSTEM OVERVIEW

The container implements `WAV path → float`. Four stages run per item:

1. **Transcription.** Transkum [1] converts the recording to a MIDI note sequence with onsets, offsets, velocities, and sustain pedal.
2. **Encoding.** The sequence is tokenised and passed through MoonBeam-839M [2], a MIDI-native causal language model, carrying LoRA adapters [3]. Sequences longer than the 1024-token context are split into consecutive chunks and every chunk is encoded; encoder cost is therefore linear in piece duration.

3. **Pooling.** Token embeddings are averaged over the whole piece.
4. **Regression.** A ridge head maps the pooled embedding to a scalar.

We chose transcription-first over an audio-native model after a frozen bake-off in which MoonBeam-839M with mean pooling reached τ_c 0.8257 against 0.7790 for a competing symbolic encoder on identical folds. A control established that the gain is backbone-specific rather than an artefact of the pooling operator.

2. TRAINING DATA

Dataset. Piano Syllabus (PSyllabus) [4, 5], Zenodo 10.5281/zenodo.14794592: 7,901 recordings of solo piano repertoire, graded on an 11-level scale aggregated from 18 grading systems. No other corpus contributed to training.

Size actually used. PSyllabus distributes labels and video identifiers rather than audio, so recordings were fetched individually and transcribed locally. We obtained and transcribed **7,665 of 7,901 pieces (97.0%)**; the remaining 236 are no longer retrievable (186 removed, 48 private, 2 region-blocked). The five submitted adapters were each trained on all 7,665, with nothing held out.

Preprocessing. Audio is transcribed to MIDI with Transkum; no manual cleaning, filtering, or repertoire curation was applied. Training samples a single random 1024-token window per piece per epoch, while inference encodes every chunk and mean-pools, so the training and inference views deliberately differ. We swept the training window length and retained 1024.

Contamination control. The task specifies that the evaluation repertoire is drawn from sources outside existing public piano difficulty datasets. PSyllabus is such a dataset, so disjointness holds by construction. We disclose the corpus in full so that organisers can verify this directly; at the time of writing the announced list of forbidden pieces and composers had not been published, so no intersection against it could be computed. An automated check is held ready to run against that list should it appear.

3. MODEL AND LABEL SPACE

Architecture. MoonBeam-839M with LoRA on the top 5 of 15 transformer layers, adapting the seven projection matrices in each ($\{q, k, v, o, \text{gate}, \text{up}, \text{down}\}$, 35 modules).



The backbone is otherwise frozen. Training ran 3 epochs at learning rate 10^{-4} .

Label space. PSyllabus grades are ordinal with 11 levels, and the published grade is the rounded mean of the judgements of one or more grading boards ($r = 0.981$ against the raw board mean). We apply *no* remapping: the head is fit by ridge regression directly on the integer grade and emits an unrounded real value, because the task’s metric is rank-based and its own scale is undisclosed. Emitting a continuous score avoids committing to a granularity we cannot observe.

Post-processing. The submitted score is the unweighted mean of the five member predictions. Members differ only by training seed (2026–2030), which varies both LoRA initialisation and window sampling. Scores are neither clipped nor rounded. On an item that fails to transcribe or encode, the container emits the corpus median and logs the failure to `stderr`; this bounds the effect of unreadable inputs on the run as a whole rather than aborting it.

4. VALIDATION

All figures below come from composer-disjoint 5-fold cross-validation over 900 evaluation pieces spanning 900 distinct composers, with byte-identical folds across arms and paired bootstrap over pieces (2000 resamples, seed 2026).

arm (identical folds, $n = 900$)	τ_c
37 hand features, ridge (matched)	0.8068
MoonBeam-839M frozen, mean-pool	0.8257
MoonBeam-839M + LoRA, single adapter	0.8430
5-adapter prediction mean (recipe)	0.8461
<i>estimated label ceiling</i>	<i>0.8668</i>

The fine-tuned encoder beats the supervision-matched hand-feature baseline by a margin whose 95% bootstrap interval excludes zero, and the five-adapter ensemble beats a single adapter by $+0.0051$ CI95[$+0.0019, +0.0082$]. The per-adapter spread (std 0.0033) is the same order as the ensemble gain, which is consistent with variance averaging rather than any member holding unique signal.

Label ceiling. Fitting a graded response model [6] to all 13,659 raw board judgements, with per-board discriminations and thresholds, gives the τ_c that a *perfect* model would score against this label: **0.8668** CI95[0.8549, 0.8805]. Total remaining headroom on this corpus is therefore about $+0.024$. A posterior-predictive check indicates the fit understates board agreement, so the estimate is conservative. This ceiling is specific to this corpus’s board composition; the MIREX test set is independently curated and its own ceiling is unmeasurable from here.

What we do not claim. Published results on PSyllabus are not comparable to the figures above: they use random rather than composer-disjoint splits, a different subset, and a different grade mix. We make no claim of improving on

them. We likewise report no τ_c for the submitted model, because every labelled piece is in its training set.

5. COMPUTE

Measured on one NVIDIA L4 over 8 recordings spanning 15–1806 s, 0/8 failures: the five-member ensemble costs $5.00\text{ s} + 0.0410 \times \text{duration}$ per item, against $4.93\text{ s} + 0.0360 \times \text{duration}$ for a single adapter. An earlier run of the same protocol on a separately built copy of the model gave $4.97\text{ s} + 0.0400 \times \text{duration}$, so the figure reproduces. At our corpus’s median piece length of 103 s this is an **average inference time of 9.2 s per piece**, or roughly 9,400 items per 24 hours on one GPU; at the mean duration of our evaluation set (150 s) the average is 11.2 s per piece. The ensemble’s cost falls almost entirely on the slope rather than the fixed term, because the encoder processes every chunk of a piece.

Peak GPU memory is 5,976 MiB (roughly 5.8 GiB), sampled device-wide at 0.5 s intervals across the whole run. We report the device figure rather than the framework’s own counters because the pipeline spans two processes: the encoder’s allocator reserved 3,858 MiB, and the remainder is CUDA context plus the Transkun transcriber, which runs in a separate process with its own context. A framework-only number would have omitted the transcriber entirely. An 8 GiB card is sufficient.

The container also requires **10 GiB of host RAM** (9 GiB is killed during model load), because the encoder and the transcriber occupy separate processes. The submitted image is self-contained and performs no network access at inference time.

6. REFERENCES

- [1] Y. Yan, “Transkun: Automatic piano transcription,” software, MIT licence. <https://github.com/Yujia-Yan/Transkun>.
- [2] Z. Guo, “MoonBeam: A MIDI foundation model,” 2025, Hugging Face model repository, Apache 2.0. <https://huggingface.co/guozixunnicolais/moonbeam-midi-foundation-model>.
- [3] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “LoRA: Low-rank adaptation of large language models,” in *International Conference on Learning Representations (ICLR)*, 2022.
- [4] P. Ramoneda, M. Lee, D. Jeong, J. J. Valero-Mas, and X. Serra, “Can audio reveal music performance difficulty? Insights from the Piano Syllabus dataset,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2025, arXiv:2403.03947.
- [5] —, “Piano Syllabus dataset,” 2025, Zenodo, doi:10.5281/zenodo.14794592.
- [6] F. Samejima, “Estimation of latent ability using a response pattern of graded scores,” *Psychometrika Monograph Supplement*, 1969.