

# RNInfer: A Large Language Model Approach to Functional Harmonic Reasoning in Symbolic Music



Hewei Gao\*<sup>1</sup>, Tim Beyer<sup>1</sup>, Xiaoxuan Wang<sup>2</sup>, Stephan Günnemann<sup>1</sup>, Martin Alois Rohrmeier<sup>2</sup>

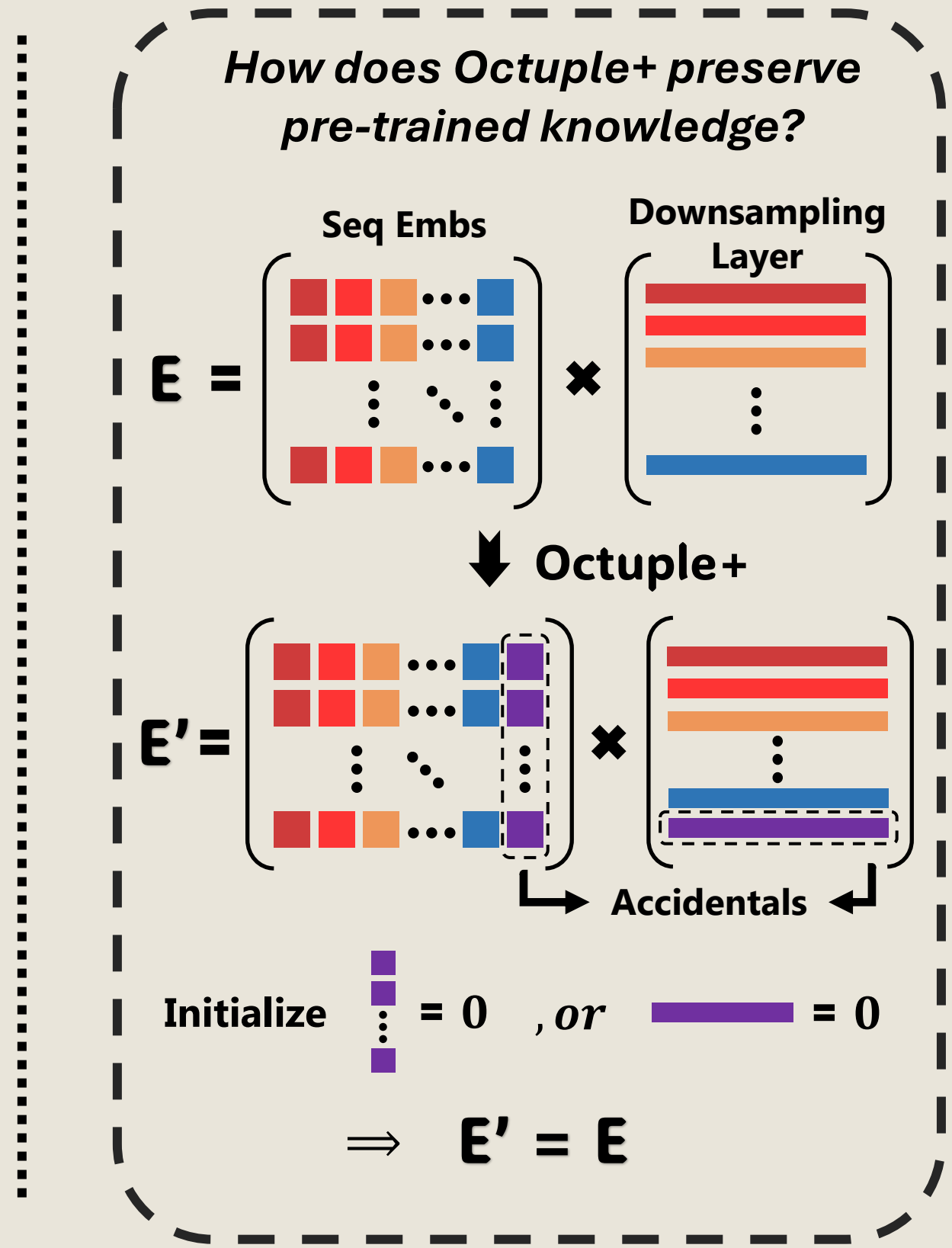
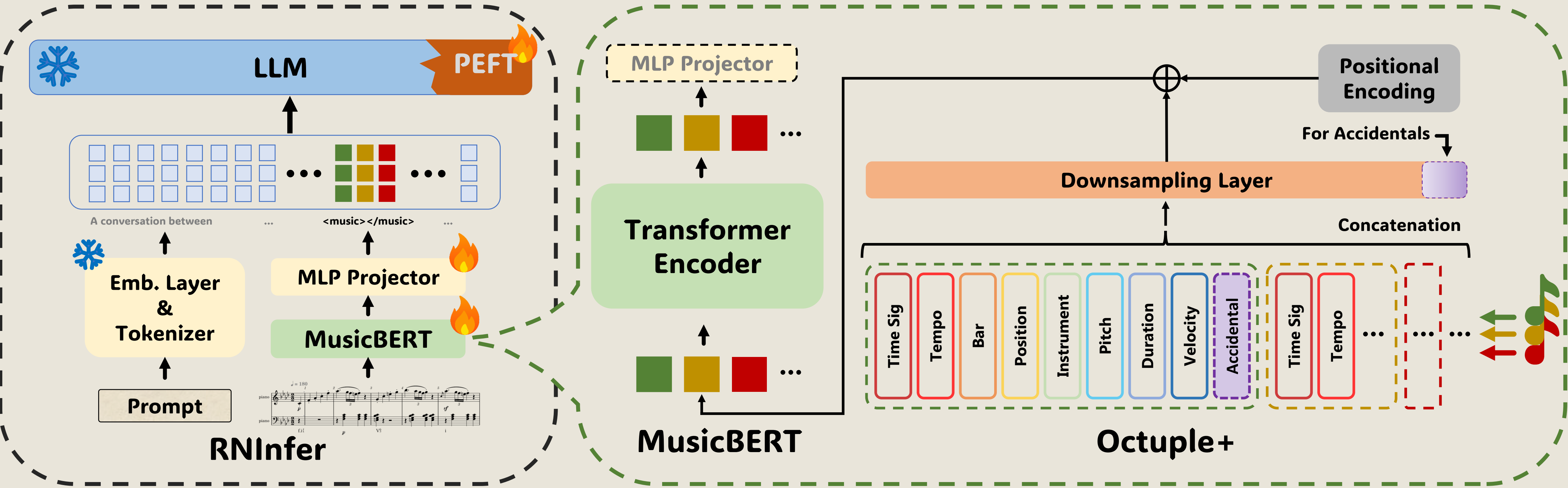
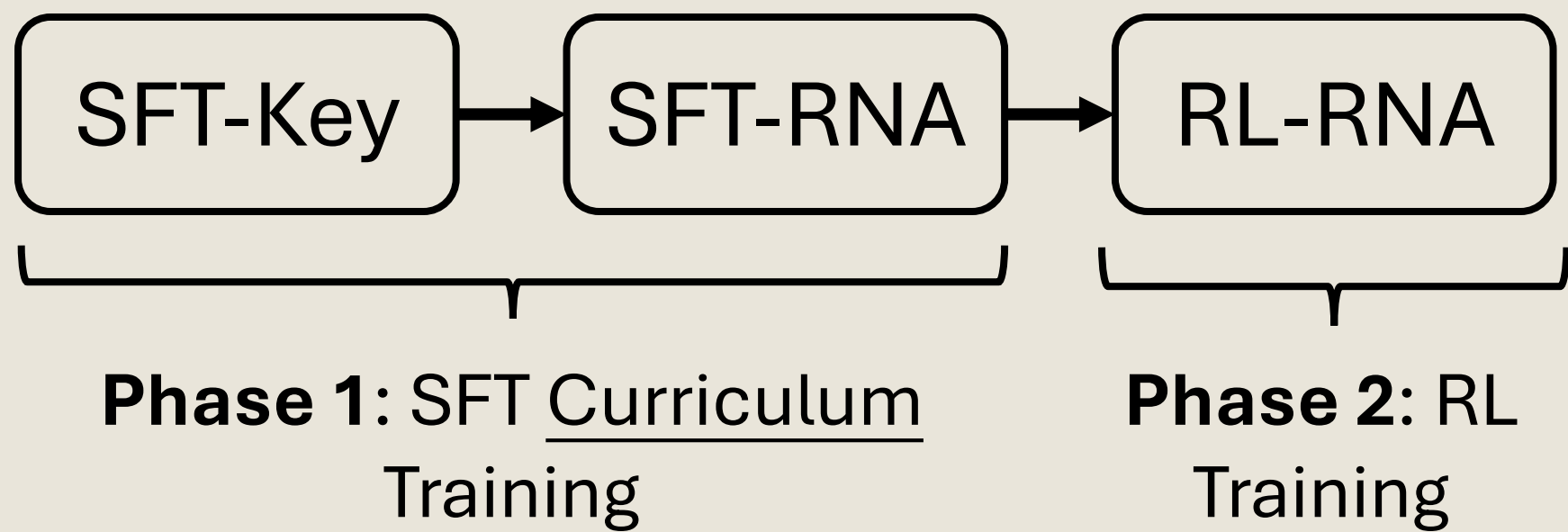
<sup>1</sup>Technical University of Munich

<sup>2</sup>École Polytechnique Fédérale de Lausanne

## TL;DR

- Roman Numeral Analysis (RNA) is a key tool for analyzing harmonic progression in Western classical music. This study is the first to apply LLMs to RNA.
- We apply RL (e.g., GRPO) to train LLMs as reasoning models, with the goals of:
  - a) validating VLM-like architectures for symbolic music analysis, and
  - b) overcoming the interpretability limitations of traditional classification models.

## Training Process



## Choice Validation

Model	Position	Group	Both
RNInfer	99.51%	99.74%	99.35%

Table 1: Validation of the MLP projector on the simultaneous-note determination task.

Tokenization	Key Acc. (%)
Octuple (w/o accidental)	77.04
Octuple+ (w/ accidental)	<b>79.46</b>

Table 2: Validation of **Octuple+** on the SFT-Key task using the RNBert encoder.

SFT-RNA Training	Acc. (%)
From Scratch	29.44
With SFT-Key curriculum	<b>49.51</b>

Table 3: Validation of **curriculum training** in the SFT phase.

Slicing Strategy	Acc. (%)
Measure-based	73.90
Phrase-based	<b>79.46</b>

Table 4: Validation of the **phrase-based slicing strategy** on the SFT-Key task.

Music Encoder	Acc. (%)
MusicBERT (pre-trained)	21.69
RNBert (fine-tuned)	<b>49.51</b>

Table 5: Validation of the **domain-specific fine-tuned music encoder** on the SFT-RNA task.

## Main Results

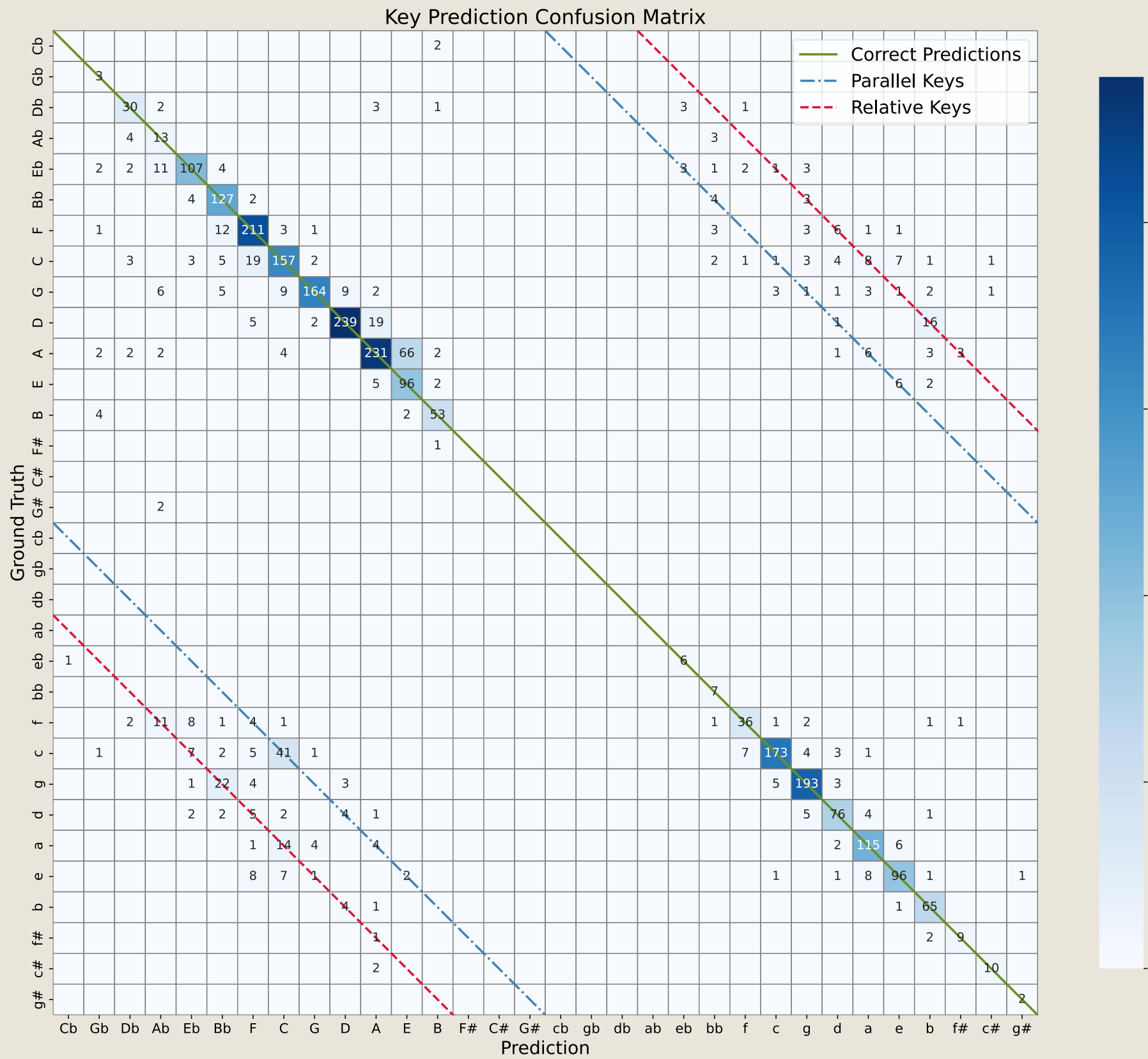
### SFT-Key

Model	Dataset	Acc. (%)
RNInfer	BPS	80.33
RNBert	BPS	<b>81.50</b>
RNInfer	P.DCML	79.36
RNBert	P.DCML	<b>85.76</b>

Table 6: SFT-Key task accuracy (%) compared to RNBert.

- **RNInfer achieves performance comparable to SOTA approaches in the key-prediction task**, despite the greater difficulty of its task setting (e.g., enharmonic keys, open-ended generation.)

- **RNInfer implicitly learns tonal relationships**, as shown by the distribution of its mispredictions in the confusion matrix:
  - Main diagonal: correct predictions
  - Right = dominant, Left = subdominant
  - Red dashed line: relative keys
  - Blue dashed line: parallel keys

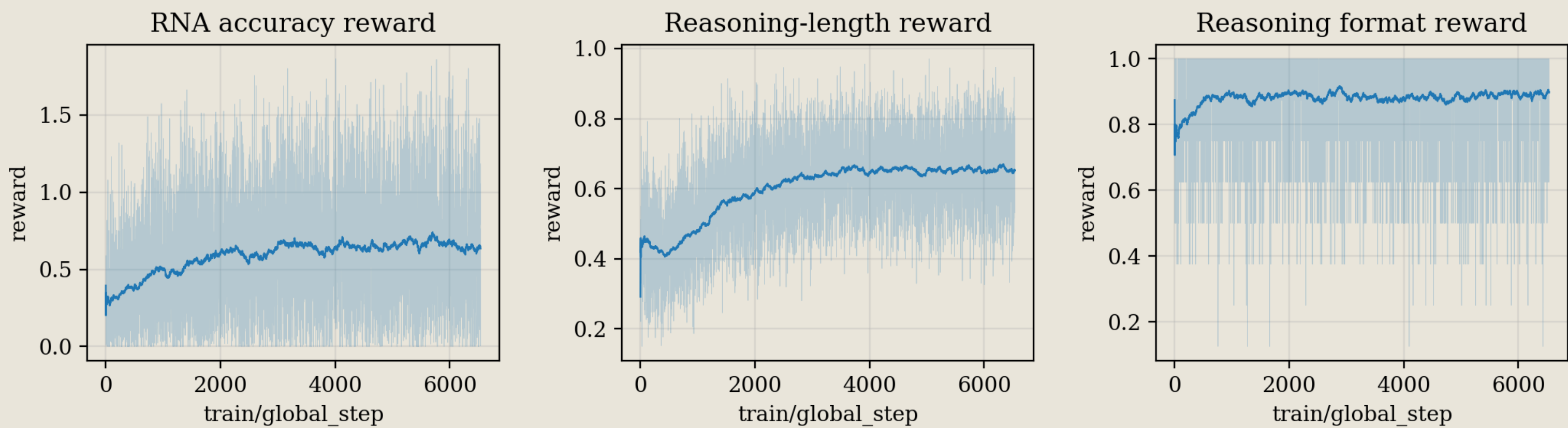


### RL-RNA

Model	Training Stage	Dataset	Acc. (%)
RNInfer	SFT-RNA	BPS	<b>58.28</b>
RNInfer	RL-RNA	BPS	38.67

Table 8: Accuracy (%) before and after RL-RNA.

- Accuracy decreases, although the average rewards for all three reward functions steadily increase during training.
- **Main Reason**: The absence of a formal verifier for reasoning-trace validity.



### Example



<think>  
**Step 1:** Establish the local key.- The local key is f. There are no significant changes in the key signature or chord progression that would indicate a shift to a different key.  
**Step 2:** Confirm/adjust the harmonic-rhythm.- The harmonic rhythm is consistent with one chord per bar.  
**Step 3:** Filter non-chord tones.- No non-chord tones are present.  
**Step 4:** Identify chord quality and state the root-position Roman numeral.- bar 0: i: F-Ab-C, R-T-F - bar 2: V65: C-E-G-Bb, R-T-F-S - bar 4: i: F-Ab-C, R-T-F - bar 5: #vii6: E-G-Bb, R-T-F - bar 6: i6: F-Ab-C, R-T-F - bar 6: iio6: G-Bb-Db, R-T-F - bar 7: V: C-E-G, R-T-F - bar 8: v: C-Eb-G, R-T-F  
**Step 5:** Determine inversion.- [same output as in Step 4]  
**Step 6:** Detect tonicizations/modulations.- No tonicizations or modulations are present.  
**Step 7:** Proof-read & summarise higher-level pattern.- The excerpt is in f, with a progression of i, V65, i, #vii6, i6, iio6, V, v.  
</think>

<answer>  
{  
 "f": {  
 "bar 0 quarterbeats 1.00: i: F-Ab-C, R-T-F",  
 "bar 2 quarterbeats 1.00: V65: C-E-G-Bb, R-T-F-S",  
 "bar 4 quarterbeats 1.00: i: F-Ab-C, R-T-F",  
 "bar 5 quarterbeats 1.00: #vii6: E-G-Bb, R-T-F",  
 "bar 6 quarterbeats 1.00: i6: F-Ab-C, R-T-F",  
 "bar 6 quarterbeats 3.00: iio6: G-Bb-Db, R-T-F",  
 "bar 7 quarterbeats 1.00: V: C-E-G, R-T-F",  
 "bar 8 quarterbeats 1.00: v: C-Eb-G, R-T-F"  
 }  
}  
</answer>

### SFT-RNA

- RNInfer outperforms strong baselines but does not yet surpass RNBert.

#### Reasons

- Open-ended generation vs. classification.
- 42 enharmonically distinct keys vs. 24 in RNBert.
- Support for more chord types than RNBert.

Model	Training set	Test set	Acc. (%)
RNInfer	P.DCML	BPS	<b>58.3</b>
AugN [2]	AugNData	BPS	45.4
Mi20 [4]	Mi20Data	BPS	42.8
RNInfer (8B)	P.DCML	P.DCML	52.8
RNInfer	P.DCML	P.DCML	49.5
RNBert	P.DCML	P.DCML	<b>64.5</b>

Table 7: Results for the SFT-RNA task. The P.DCML dataset is comparable in size to those used for AugN and Mi20.