# TTR-2: a native Text-to-Ringtone model

*Where the Text-to-Ringtone category goes next. A research agenda for training a generative model end-to-end on the ringtone form itself — and aligning it with reinforcement learning from how people actually use ringtones.*

Position paper · Ringoz Research, YESH Studios LLC · Published 2026-07-07 · Status: research agenda — this paper describes proposed future work, not the shipping product.

## Toward TTR-2: learning the ringtone form end-to-end with Reinforcement Learning from Ringtone Feedback

### Abstract

The ringtone is one of the most frequently heard audio artifacts in the world and one of the least studied generative targets. Today, Ringoz serves the Text-to-Ringtone category with TTR-1, a proprietary multi-stage engine we engineered for the 30-second ringtone form. This paper describes what we intend to build next: TTR-2, a single generative model trained natively and end-to-end on ringtones themselves. We propose (i) a training corpus on the order of hundreds of thousands of prompt–ringtone pairs, assembled from licensed productions, commissioned recordings, and curated synthetic exemplars; (ii) a discrete audio-token language model conditioned on name, phrase, genre, and language; and (iii) a final alignment stage we call **Reinforcement Learning from Ringtone Feedback (RLRF)**, in which the model is optimized against

reward models for prompt adherence, sung-name intelligibility, loop continuity, small-speaker salience, and human musical preference — the last grounded in the strongest signal in this domain: whether a person actually sets the result as their ringtone. We further propose RingBench, an open evaluation protocol for the text-to-ringtone task. This is a position paper: it defines our research north star, the reasons we believe the ringtone deserves its own model, and the standards we will hold that model to.

# 1. Introduction

Generative audio research has converged on two well-served targets: speech and full-length music. The ringtone sits between them and belongs to neither. It is musical but not a song; it carries words but is not speech; it is heard more often than almost any other audio a person owns, under the worst playback conditions consumer audio has to offer — a small speaker, in a pocket, in a room full of other sound.

Ringoz introduced the Text-to-Ringtone (TTR) category: a short text prompt — a name, a phrase, a vibe — becomes a fully produced, loop-ready 30-second ringtone with sung vocals. The shipping engine, TTR-1, is a multi-stage generative pipeline we engineered specifically for this form; its stage architecture is described publicly at /technology and its implementation is proprietary. TTR-1 is the present. This paper is about the destination.

Our thesis is simple: **a form this constrained, this repeated, and this measurable deserves a model whose only training objective is the form itself.** General music models optimize for plausible music under a text description. A ringtone-native model must optimize for something stricter — a hook that lands in the first seconds, a name that survives a phone speaker, an ending that loops seamlessly into its own beginning, and a listener who, given the choice, keeps it. Every one of those properties can be measured. Measurable properties can be turned into rewards. That is why the second half of this agenda is reinforcement learning.

# 2. The ringtone as a generative form

We define the ringtone-native generation task by five properties that jointly distinguish it from general text-to-music:

- **Identity payload.** The prompt usually contains a proper name or personal phrase, and the artifact fails if that payload is not clearly intelligible when sung. Intelligibility of an arbitrary name — across 12 vocal languages — is a hard constraint, not a stylistic preference.
- **Hook-first temporal budget.** A phone may be answered within two or three seconds. The musical payoff cannot be built toward; it must be front-loaded, then sustain interest across the remaining window.
- **Loop closure.** Ringtones repeat. The final instant must hand off to the first without an audible seam — a boundary-continuity condition that full-length music generation never has to satisfy.
- **Playback salience.** The target device is a phone speaker, often occluded. Spectral balance, loudness, and transient design must survive a transfer function that destroys most produced music.
- **Bounded length.** The form is 30 seconds. This is a constraint, but also a gift: short sequences mean cheaper training, denser feedback, and far more optimization steps per unit of compute than full-length music affords.

None of these properties is optimized by models trained to continue or describe general music. All of them are objectively checkable, which makes the ringtone — perhaps uniquely among creative audio forms — a natural reinforcement-learning environment.

# 3. Related work

**Text-to-music generation.** Autoregressive modeling over discrete audio tokens established that language-model architectures transfer to audio: Jukebox [3] generated raw-audio music with lyrics, AudioLM [13] framed audio continuation as language modeling, and MusicLM [1] and MusicGen [2] made text-conditioned music generation practical at quality. These systems target open-ended music; none optimizes for loop closure, name intelligibility, or small-speaker playback.

**Discrete audio representations.** Neural codecs with residual vector quantization — SoundStream [4] and EnCodec [5] — compress waveforms into token sequences that

language models can ingest and emit. The 30-second bound of the ringtone form keeps these sequences short, which we regard as a structural advantage of the task.

**Singing voice synthesis.** Work such as DiffSinger [10] shows that sung vocals with controllable lyrics are learnable. The ringtone setting is harder in one respect — arbitrary proper names as lyric content, in 12 languages — and easier in another: the lyric is short.

**Preference alignment.** Reinforcement learning from human feedback aligned language models with what people actually want [6], with PPO [7] and later direct preference optimization [8] and group-relative methods [12] as the workhorse algorithms, and AI feedback extending the paradigm where human labels are scarce [9, 11]. MusicRL [14] demonstrated that music generation specifically benefits from preference fine-tuning at scale. Our proposal takes the next step the ringtone uniquely enables: a domain where the preference signal is not a rating collected in a lab, but a decision users make anyway — set it, keep it, or regenerate.

To our knowledge, no published work treats the ringtone as a native generative target. That is the gap this research program occupies.

# 4. Problem formulation

Let the conditioning context be $c$ = (name, phrase, genre, language), and let $a$ denote a 30-second audio artifact. TTR-2 is a generative policy $\pi_\theta(a \mid c)$ over discrete audio tokens. Purely supervised training maximizes likelihood over a corpus of $(c, a)$ pairs, which teaches the distribution of good ringtones but not the boundaries of the form: likelihood does not penalize an audible loop seam or a garbled name, because both can be locally probable. The properties in Section 2 are perceptual, global, and mostly non-differentiable. We therefore formulate the final training stage as reward maximization:

$$\text{maximize} \;\; E_c \, E_{a \sim \pi_\theta} [\, R(a, c) \,] \;-\; \beta \cdot KL(\, \pi_\theta \,\|\, \pi_{SFT} \,)$$

where the KL anchor to the supervised policy preserves musicality and diversity while the reward $R$ enforces the form. The reward decomposes into the five measurable properties of Section 2, detailed in Section 7.

# 5. The corpus: data at ringtone scale

The training corpus is the center of gravity of this program. We will assemble a paired prompt–ringtone corpus on the order of **hundreds of thousands of exemplars**, built from three complementary sources:

- **Licensed and commissioned productions.** Professionally produced short-form musical pieces, licensed for training or commissioned directly, with sung vocal stems and per-piece metadata (genre, tempo, key, vocal language, hook position).
- **Curated synthetic exemplars.** Outputs of our production engine that passed both automated quality gates and real-world acceptance, used as exemplars of the target distribution under a strict curation policy: synthetic data enters the corpus only when it satisfies the same objective form-checks (Section 7) that will later serve as rewards, so the model never trains on artifacts it will be penalized for producing.
- **Structured augmentation.** Controlled recombination of licensed stems — vocal hooks re-rendered over alternative genre beds, transpositions, tempo variants — to cover the long tail of the 12-genre × 12-language grid where organic data is thin.

Separately from the audio corpus, production telemetry provides the preference signal for alignment: aggregate, anonymized outcome events such as preview completion, regeneration, and — the strongest label in this domain — whether a generated ringtone was set and kept. We already publish aggregate style-level analyses of this telemetry in our Ringtone Trends research; RLRF will consume the same class of signal at the level of individual anonymized outcomes. No prompts, names, or personal content are ever part of published research, and Section 9 states the rights and privacy commitments the corpus will be held to.

# 6. Model architecture

We propose a two-component architecture, deliberately conventional at the component level so the novelty concentrates where the ringtone demands it:

- **A neural audio codec** with residual vector quantization, trained (or fine-tuned) with the small-speaker transfer function in its perceptual loss, so that the token space itself favors sounds that survive the target playback channel.
- **An autoregressive transformer** over the codec's discrete tokens, conditioned on the encoded context $c$. Two form-specific mechanisms distinguish it from a generic music LM: a **loop-aware decoding scheme** in which the sequence is modeled with circular

boundary context, so the final tokens are generated in view of the opening tokens they must resolve into; and **lyric-position conditioning** that binds the name tokens of the prompt to an explicit hook window, giving the intelligibility reward (Section 7) a handle the model can learn to satisfy.

The 30-second bound keeps token sequences short enough that both mechanisms are tractable at training time rather than inference-time patches.

# 7. Training program

STAGE A

### Audio-language pretraining

Pretrain the token LM on broad licensed musical audio for general musical competence — harmony, rhythm, instrumentation, sung phrasing. Nothing ringtone-specific yet; this stage buys the prior that keeps later stages musical.

STAGE B

### Supervised fine-tuning on the form

Fine-tune on the prompt–ringtone corpus (Section 5). The model learns the shape of a ringtone — hook-first structure, 30-second arcs, genre-conditioned vocal delivery across 12 languages — as a distribution to imitate.

STAGE C

### RLRF — Reinforcement Learning from Ringtone Feedback

Optimize the policy against the composite ringtone reward with a KL anchor to Stage B. On-policy methods (PPO-family, group-relative baselines) where rewards are cheap to query; offline preference optimization on paired user outcomes where they are not.

The composite reward instantiates Section 2, one term per property:

| REWARD TERM | MEASURES | GROUNDING |
|---|---|---|
| $\mathbf{R_{adhere}}$ | Prompt, genre, and language adherence | Joint text–audio embedding similarity, per-genre classifier heads |
| $\mathbf{R_{name}}$ | Sung-name intelligibility | Lyric transcription of the generated vocal aligned against the prompt name; word-error penalty |
| $\mathbf{R_{loop}}$ | Loop closure | Spectral and energy continuity across the end→start boundary; audible-seam detector |
| $\mathbf{R_{salience}}$ | Small-speaker playback survival | Psychoacoustic loudness and spectral balance measured through a phone-speaker transfer function |
| $\mathbf{R_{pref}}$ | Human musical preference | Reward model trained on rater comparisons plus aggregate anonymized set/keep/regenerate outcomes |

Composite reward hacking is the known failure mode of this design — a policy that maximizes the measurable terms while degrading the experience they proxy. Our mitigations are standard but non-negotiable: the KL anchor to the supervised policy, held-out reward-model ensembles that are never optimized against directly, hard constraint thresholds on $R_{name}$ and $R_{loop}$ rather than pure scalarization, and continuous human audit sampling of the policy's outputs throughout training.

# 8. Evaluation: RingBench

A model trained on a new task needs a benchmark for that task. We propose **RingBench**, a fixed evaluation protocol we intend to publish alongside the model's results:

- A frozen prompt panel spanning the 12-genre × 12-language grid, including adversarially hard names (rare phonologies, cross-language names, very short and very long inputs).
- **Objective metrics**: name word-error rate from lyric transcription; loop-boundary discontinuity; loudness and spectral-balance deltas through the phone-speaker transfer function; text–audio adherence score.
- **Human evaluation**: blinded pairwise preference against both the current production engine and general-purpose music generators given equivalent prompts, with "would you set this as your ringtone?" as the primary question.
- **The end metric**: set-and-keep rate under a production A/B — the only number that fully closes the loop between the research objective and the artifact's actual job.

# 9. Rights, safety, and provenance

The corpus and the model will be held to commitments we consider preconditions, not features:

- **Licensed data only.** Every training recording is licensed for training or commissioned; every vocal performance is from a consenting, compensated performer.
- **No artist imitation.** The model will not be conditioned to reproduce the voice or style of identifiable artists, and RingBench includes imitation probes as a release gate.
- **Privacy.** Preference telemetry is aggregate and anonymized; prompts and names never leave the production boundary, and no personal content enters published research.
- **Content policy.** Name and phrase inputs pass the same content policy in research as in production.

# 10. Roadmap

| PHASE | OBJECTIVE | EXIT CRITERION |
|---|---|---|
| **I — Foundations** | Corpus assembly and licensing; reward-model suite built and validated against human judgment | Reward models agree with blinded raters at target reliability on held-out audio |
| **II — Supervised prototype** | Stage A + B model competitive on RingBench objective metrics | Prototype matches production output quality on adherence and form metrics |
| **III — RLRF at scale** | Stage C alignment on the full reward suite | Blinded human preference over the supervised prototype, with no regression on constraint terms |
| **IV — Production candidate** | Distillation and serving work; production A/B | Set-and-keep rate wins its A/B; RingBench results published |

We deliberately publish phases and exit criteria rather than dates. The program advances when the criteria are met.

# 11. Conclusion

The ringtone has spent two decades as an afterthought of other audio pipelines — trimmed from songs, converted between formats, chosen from catalogs. The Text-to-Ringtone category changed what a ringtone is; TTR-2 is our commitment to changing how one is made at the deepest level available: a model that has never known any other objective than this form, trained on hundreds of thousands of examples of it, and aligned by reinforcement learning against the only judgment that matters — the person whose phone is about to ring.

We publish this agenda as our north star. Progress against it will appear here, on the changelog, and in future Ringoz Research releases.

# References


1. Agostinelli, A. et al. (2023). MusicLM: Generating Music From Text. arXiv:2301.11325.
2. Copet, J. et al. (2023). Simple and Controllable Music Generation. arXiv:2306.05284.
3. Dhariwal, P. et al. (2020). Jukebox: A Generative Model for Music. arXiv:2005.00341.
4. Zeghidour, N. et al. (2021). SoundStream: An End-to-End Neural Audio Codec. arXiv:2107.03312.
5. Défossez, A. et al. (2022). High Fidelity Neural Audio Compression. arXiv:2210.13438.
6. Ouyang, L. et al. (2022). Training Language Models to Follow Instructions with Human Feedback. arXiv:2203.02155.
7. Schulman, J. et al. (2017). Proximal Policy Optimization Algorithms. arXiv:1707.06347.
8. Rafailov, R. et al. (2023). Direct Preference Optimization: Your Language Model is Secretly a Reward Model. arXiv:2305.18290.
9. Bai, Y. et al. (2022). Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073.
10. Liu, J. et al. (2022). DiffSinger: Singing Voice Synthesis via Shallow Diffusion Mechanism. arXiv:2105.02446.
11. Lee, H. et al. (2023). RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. arXiv:2309.00267.
12. Shao, Z. et al. (2024). DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv:2402.03300.
13. Borsos, Z. et al. (2022). AudioLM: A Language Modeling Approach to Audio Generation. arXiv:2209.03143.
14. Cideron, G. et al. (2024). MusicRL: Aligning Music Generation to Human Preferences. arXiv:2402.04229.


# About Ringoz

Ringoz is the first Text-to-Ringtone engine. Its TTR-1 pipeline converts a short text prompt — a name, a phrase, a vibe — into a fully produced, loop-ready 30-second

ringtone with sung vocals, across 12 music genres and 12 languages. Available on iPhone and Android.

Questions about this research agenda: contact@ringoz.ai. How the shipping engine works today: the TTR-1 engine. We don't discuss TTR-1's internals.