# Pronouncing Hard Words: Phoneme adaptation of a Pretrained Large Language Model-Based Speech Synthesis Model

Anonymous ACL submission

## Abstract

Modern end-to-end text-to-speech (TTS) systems produce natural and expressive speech, but explicit pronunciation control remains difficult for out-of-vocabulary words, named entities, and other pronunciation-sensitive spans. We study a general adaptation recipe for adding phoneme controllability to a pretrained LLM-based TTS model without retraining from scratch. The recipe extends the text vocabulary with ARPAbet special tokens, performs continued supervised fine-tuning on phoneme-augmented data, and applies pronunciation-focused direct preference optimization (DPO) with a frame mask centered on the forced-aligned target entity. The adapted model supports both phoneme-only prompts and hybrid prompts in which local phoneme spans are embedded inside ordinary text. On held-out drug names excluded from DPO preference construction, supervised phoneme adaptation achieves a 20.96% entity-span phoneme error rate, and DPO with SFT replay improves this to 11.53%. The final adapted model also outperforms the evaluated open- and closed-source TTS baselines on the drug-pronunciation benchmark. Beyond pronunciation accuracy, Seed-TTS-Eval performance is preserved, but DPO can shift speaking rate and weaken expressive control, motivating evaluation of phoneme adaptation as a multi-objective problem.

## 1 Introduction

Recent text-to-speech (TTS) systems have moved beyond clean single-speaker narration toward expressive, controllable, and conversational speech generation. Large language model and diffusion or flow-matching based architectures have improved naturalness, prosody, zero-shot voice cloning, and streaming synthesis in systems such as Qwen3-TTS (Hu et al., 2026) and CosyVoice 3 (Du et al., 2025). In parallel, dialogue-oriented speech generation systems such as FireRedTTS-2 (Xie et al., 2025), CoVoMix (Zhang et al., 2024), CoVoMix2 (Zhang et al., 2025a), and MOSS-TTSD (Zhang et al., 2026) have pushed the field toward multi-speaker generation from dialogue scripts, context-aware turn modeling, expressive delivery, and long-form conversational stability. These capabilities make modern TTS systems increasingly attractive as the speech layer for voice agents, chat assistants, and interactive domain-specific applications.

Despite these advances, explicit pronunciation control has received comparatively less attention than expressivity, speaker cloning, and dialogue modeling. This omission becomes important as voice agents are deployed in specialized domains such as healthcare, legal services, and finance, where they encounter rare, trust-critical entities whose pronunciations are not reliably recoverable from spelling alone. Drug names such as *Mounjaro* and *Xiidra*, legal terms such as *voir dire* and *certiorari*, and brand names such as *Kweichow Moutai* and *Moet Hennessy Louis Vuitton* are short spans, but mispronouncing them can dominate the perceived quality of an interaction. These examples motivate the need for explicit pronunciation control across specialized domains.

Classical TTS pipelines exposed pronunciation through front-end components such as lexicons and grapheme-to-phoneme conversion. End-to-end models simplified this pipeline by learning from text and speech directly, but in doing so they often made grapheme-to-sound decisions implicit, as seen in Tacotron (Wang et al., 2017), Tacotron 2 (Shen et al., 2018), and Deep Voice 3 (Ping et al., 2018). Prior work has therefore explored mixed grapheme and phoneme inputs through Representation Mixing (Kastner et al., 2019), phoneme-aware encoders such as PnG BERT (Jia et al., 2021), neural lexicon readers (He et al., 2022), and post-hoc correction systems such as Speech Audio Corrector (Fong et al., 2022) and Spell4TTS (Fong et al.,

2023). These results show that explicit phonetic information remains useful, but they do not fully answer a practical question raised by current pretrained dialogue TTS systems: can an existing large TTS model be adapted to accept phoneme control without retraining it from scratch?

This question matters because the best contemporary systems are trained at a scale that most teams cannot reproduce. FireRedTTS-2, for example, builds its foundational TTS ability from 1.1 million hours of monologue speech and then post-trains on 300 thousand hours of multi-speaker dialogue data (Xie et al., 2025). Qwen3-TTS reports an even larger multilingual training mixture, with over 5 million hours of speech spanning 10 languages (Hu et al., 2026).

Our goal is both methodological and practical. We describe the training choices needed to make phoneme input work in a pretrained dialogue TTS model, the evaluation protocol needed to measure localized pronunciation control, and the pitfalls that arise when automatic phoneme recognizers are used to construct preference data. Our results show that pronunciation-focused DPO improves corpus-level recognition metrics over supervised fine-tuning, but also expose a recognizer-derived bias toward slower speech. This trade-off motivates our analysis of not only whether phoneme adaptation works, but also what kinds of artifacts it can introduce. Our contributions include:

- We show how to adapt an open-source pretrained TTS model for hybrid text-ARPAbet phoneme input.
- We treat phoneme annotation rate as a design knob, sweep it across seven values from 0 % to 100 %, and identify an operating point that balances pharmaceutical-entity pronunciation against general synthesis quality.
- We introduce a pronunciation-focused DPO objective that compares candidates on a forced-aligned acoustic region centered on the target entity span, with an SFT-replay term that preserves expressivity.
- We introduce an inference-time reference-biasing mechanism that conditions on a preceding-turn pronunciation reference, raising the mean drug-pronunciation rate and tightening 5-run variance with no parameter updates.
- We analyze side effects introduced by pronunciation-focused DPO, including speaking-rate shifts and emotion-tag steering loss, and propose measurements and mitigation strategies for detecting and counteracting such preference-induced artifacts.

## 2 Related Work

End-to-end systems such as Tacotron (Wang et al., 2017; Shen et al., 2018) and Deep Voice 3 (Ping et al., 2018) showed that high-quality speech can be synthesized directly from text, but they made grapheme-to-sound decisions implicit, leaving users few ways to specify how rare names, domain entities, or out-of-vocabulary words should be pronounced. A first line of work retains partial phonetic control: Representation Mixing (Kastner et al., 2019) combines character and phoneme inputs in a single encoder, Fong et al. (2020) show that even limited phonemic supervision improves controllability, and Deep Voice 3 also supports mixed character/phoneme inputs. A complementary line corrects pronunciation post-hoc rather than rebuilding the front-end, as in Speech Audio Corrector (Fong et al., 2022) and Spell4TTS (Fong et al., 2023). A third line enriches the text representation itself: Neural Lexicon Reader (He et al., 2022) extracts pronunciation from external textual sources, and PnG BERT (Jia et al., 2021), Mixed-Phoneme BERT (Zhang et al., 2022), and XPhoneBERT (Nguyen et al., 2023a) introduce phoneme-aware contextual representations. Production front-ends still treat G2P, homograph disambiguation, and post-lexical rules as active modeling problems (Comini et al., 2023). These works establish that phonetic information matters, but they primarily modify encoders, front-ends, or correction modules.

Contemporary open-source TTS systems expose uneven phoneme-control support. FastPitch (Łańcucki, 2021) is commonly deployed with ARPAbet front-ends, while CosyVoice 3 (Du et al., 2025) and VoxCPM (Zhou et al., 2025) emphasize scalable generation, voice cloning, and controllability with optional inline phonetic paths. StyleTTS 2 (Li et al., 2023) can be paired with external phoneme or IPA front-ends, and DiaMoE-TTS (Chen et al., 2025) standardizes inputs through IPA for dialectal transfer. Zero-shot and dialogue-oriented systems such as XTTS (Casanova et al., 2024) and MOSS-TTSD (Zhang et al., 2026) primarily target multilingual cloning or dialogue gen-

eration rather than ARPAbet control. This unevenness motivates our focus on adaptation rather than model selection: we retrofit phoneme special tokens, hybrid grapheme–phoneme prompting, and localized preference optimization into a pretrained decoder-only dialogue TTS model with minimal architectural change.

Preference alignment offers a natural mechanism for repairing residual pronunciation errors after SFT. DPO (Rafailov et al., 2023) aligns generative models from pairwise preferences without reward-model RL, and recent TTS work applies preference alignment to intelligibility, speaker similarity, and subjective quality (Tian et al., 2024), as well as to difficult zero-shot settings such as repeated words, code-switching, and cross-lingual synthesis (Zhang et al., 2025b). We specialize the alignment signal to phoneme-conditioned pronunciation correctness and restrict the comparison to a forced-aligned entity span, making DPO a mechanism for stabilizing generation-to-generation pronunciation rather than a generic quality-improvement stage.

## 3 Method

### 3.1 Base Model and Input Representation

We start from the post-trained FireRedTTS-2 checkpoint (Apache 2.0 License), a long-form, multi-speaker speech synthesis model built around a streaming discrete speech tokenizer and a dual-transformer autoregressive decoder (Xie et al., 2025). FireRedTTS-2 represents speech as low-rate acoustic frames and uses a large decoder-only transformer to predict the first acoustic codebook, followed by a smaller depth decoder that fills in the remaining codebooks. This design supports long conversational contexts while retaining the codec-level structure needed for high-fidelity synthesis.

Each time step is represented internally as a 17-dimensional frame. The first 16 coordinates correspond to acoustic codebooks, while the 17th coordinate is reserved for the text token stream. The audio end-of-segment marker is represented as the all-zero frame. This factorization gives a natural locus for phoneme adaptation: we leave the acoustic tokenizer and decoders unchanged, and modify only the text-side interface.

### 3.2 Stage 1: Supervised Phoneme Adaptation

We adapt the text interface by adding 131 special tokens to the tokenizer: an extended ARPAbet phone inventory and two boundary markers, `<|arpa_start|>` and `<|arpa_end|>`. The original tokenizer contains 151,665 text tokens, while the FireRedTTS-2 text embedding table has capacity for 151,936 entries. The expanded tokenizer therefore remains within the existing embedding table, allowing us to assign the new ARPAbet symbols to unused indices without resizing the model.

We then continue supervised fine-tuning from the post-trained checkpoint. Each example keeps the same interleaved text-audio format used by FireRedTTS-2, but the text span may contain ordinary grapheme text, a fully phonemic ARPAbet (Jurafsky and Martin, 2000) transcription, or a hybrid mixture in which selected words are probabilistically replaced by their phoneme sequence inside the `<|arpa_start|>`/`<|arpa_end|>` boundaries. The ARPAbet sequence for each replaced word is looked up on the fly in the CMU Pronouncing Dictionary (Carnegie Mellon University Speech Group), which provides ARPAbet transcriptions with lexical stress for over 134,000 English entries. The replacement probability is controlled by a hyperparameter, *phoneme_tagging_rate* $\in [0, 1]$. This teaches the model to synthesize from explicit phoneme sequences while retaining the ordinary text pathway when no phoneme override is supplied; we retain plain-text examples alongside phoneme-augmented examples throughout SFT to reduce regressions in standard text-conditioned synthesis.

Alongside the phoneme augmentation, we normalize paralinguistic and emotion tags in the SFT corpora so the model learns to use FireRedTTS-2's native tag interface. The source corpora encode these cues in three incompatible ways: per-turn metadata fields in UltraVoice and Expresso, Fisher-style inline brackets, and parenthesized audio tags in the in-house data — roughly 170 surface variants in total. We map every variant onto FireRedTTS-2's existing emotion and paralinguistic special tokens (`<|laugh|>`, `<|sigh|>`, `<|breath|>`, `<|emotion_happy|>`, etc.), so no new vocabulary is added and tag→acoustic behaviour is learned in the same continued-SFT pass as the phoneme interface.

We optimize the same dual-transformer objective used by FireRedTTS-2, with an additional emphasis on the newly introduced text tokens:

$$\mathcal{L}_{\text{SFT}} = 2\Big((1-\lambda_{\text{dec}})\mathcal{L}_{\text{bb}} + \lambda_{\text{dec}}\mathcal{L}_{\text{dec}}\Big) + \lambda_{\text{text}}\mathcal{L}_{\text{text}}. \quad (1)$$

Here, $\mathcal{L}_{\text{bb}}$ is the cross-entropy loss for the backbone transformer over the first-codebook audio tokens, $\mathcal{L}_{\text{dec}}$ is the cross-entropy loss for the depth decoder over the remaining acoustic codebooks, and $\mathcal{L}_{\text{text}}$ is a text-position cross-entropy term. Following FireRedTTS-2, we compute the loss of the decoder in a randomly sampled one-eighth subset of audio frames for efficiency (Xie et al., 2025). In our phoneme adaptation stage, $\mathcal{L}_{\text{text}}$ is applied to the positions containing the newly introduced ARPAbet and special boundary tokens so that these previously unused text-token slots acquire useful representations in the text backbone. We set $\lambda_{\text{dec}} = 0.6$ and $\lambda_{\text{text}} = 0.1$.

### 3.3 Stage 2: Pronunciation-Focused DPO

SFT gives the model a baseline ability to synthesize from ARPAbet and hybrid text–phoneme input, but it does not fully stabilize pronunciation across independent generations. Even at low sampling temperatures, repeated generations for the same phoneme-controlled prompt can differ in whether the target entity is pronounced correctly. We therefore use DPO to further bias the model toward pronunciations that better match the intended phoneme sequence.

We compare two DPO masking strategies. The first is an utterance-level objective that uses all audio positions when computing chosen and rejected log-probabilities. The second is a localized objective with a windowed frame mask centered on the aligned entity span: the aligned entity frames receive the largest weight, a narrow context window around the entity receives a smaller weight, and unrelated audio frames receive zero or near-zero weight. The same full-sequence frame mask is used for both policy and reference log-probabilities, ensuring that the DPO comparison is made over the same acoustic region. We ablate this choice in Section 5.3. Let $M_e(t)$ denote the frame-weight mask for entity $e$. We compute masked log-probabilities as

$$\log \pi(y \mid x; M_e) = -\frac{\sum_t M_e(t)\, \ell_t(x, y)}{\sum_t M_e(t)},$$

where $\ell_t$ is the per-frame audio cross-entropy term.

Given a prompt $x$, a chosen sample $y^+$, and a rejected sample $y^-$, the masked DPO loss is

$$\mathcal{L}_{\text{DPO}} = -\log \sigma\Big(\beta\big[\Delta_\pi(x, y^+, y^-; M_e) - \Delta_{\pi_{\text{ref}}}(x, y^+, y^-; M_e)\big]\Big), \tag{2}$$

where

$$\Delta_\pi(x, y^+, y^-; M_e) = \log \pi(y^+ \mid x; M_e) - \log \pi(y^- \mid x; M_e). \tag{3}$$

We use the SFT model as the frozen reference policy $\pi_{\text{ref}}$ and set $\beta = 0.5$. To reduce drift in speech quality and expressivity, we add a weighted SFT replay term on supervised text-audio examples:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{DPO}} + \lambda_{\text{replay}} \mathcal{L}_{\text{SFT-replay}}. \tag{4}$$

Here, $\mathcal{L}_{\text{SFT-replay}}$ is the same cross-entropy objective used in supervised phoneme adaptation.

### 3.4 Pronunciation Biasing during Inference

To additionally stabilize the pronunciation of known difficult entities, we introduce a lightweight inference-time intervention: we teacher-force the entity's text representation and the corresponding audio tokens as a *preceding speaker turn*. This requires no parameter updates and is enabled by the multi-speaker dialogue training already present in FireRedTTS-2. The mechanism treats the reference pronunciation as ordinary conversational context, and the model reproduces it consistently in the subsequent generation. We evaluate this technique in Section 5.2.

## 4 Experimental Setup

### 4.1 SFT Data

The continued-SFT mixture preserves the pretrained model's broad synthesis ability while increasing coverage of pronunciation- and expressivity-sensitive speech. It combines four public corpora, each chosen for what it adds — Fisher English Training Speech (Cieri et al., 2004, 2005) for natural two-speaker conversation, HiFiTTS-2 (Langman et al., 2025) for clean read speech, and UltraVoice (Tu et al., 2025) and Expresso (Nguyen et al., 2023b) for paralinguistic and emotion-labeled material that supplies supervision for FireRedTTS-2's tag interface — together with an internal corpus of anonymized conversational data for domain style. The combined pool accounts for ∼50,000 hours of speech.

### 4.2 DPO Data

We construct DPO preference pairs from an internal catalog of 300 pharmaceutical-entity names paired with 50 clinical-conversational sentence templates; 270 names feed preference construction and the

remaining 30 form the held-out DRUG-OOD split. For each (name, template) pair we replace the drug surface form with its CMU-dictionary ARPAbet sequence (Section 3.2) and sample 24 candidates from the SFT model by varying random seed, temperature, and top-$k$.

Each candidate is aligned to the reference transcript with NeMo Forced Aligner (NVIDIA, 2025) to recover the entity span, and the aligned audio is decoded with a wav2vec 2.0 phoneme recognizer (Xu et al., 2021) to give an entity-span PER. Candidates with PER $< 5\,\%$ enter the chosen pool and those with PER $> 20\,\%$ enter the rejected pool; we further filter chosen candidates that exhibit consonant substitutions (e.g., *AugDyro* for *AugTyro*), which preserve low edit distance but are perceptually misleading. Since both pools are drawn from the SFT policy, the data remain on-policy and the distribution-shift issues common to off-policy DPO are reduced (Ren and Sutherland, 2024).

### 4.3 Training Details

**SFT.** Continued fine-tuning initializes from the post-trained FireRedTTS-2 checkpoint and runs for up to 30,000 steps (capped at 50 epochs) on 8 NVIDIA H200 GPUs. Optimization uses fused AdamW in bf16 with weight decay 0.01, peak learning rate $5 \times 10^{-6}$, cosine decay, and a 500-step warmup. The effective global batch size is 256.

**DPO.** DPO initializes from the selected SFT checkpoint and runs for 1,000 steps on a single NVIDIA RTX PRO 6000 GPU. We use fused AdamW in bf16 with peak learning rate $5 \times 10^{-7}$, cosine decay, and a 250-step warmup; effective batch size is 16; gradients are clipped at norm 1.0; and log-probabilities are length-normalized. The DPO temperature is set to $\beta = 0.5$, and the SFT-replay weight is $\lambda_{\text{replay}} = 1.0$ in Eq. 4.

### 4.4 Evaluation

Our evaluation targets three orthogonal questions — pronunciation accuracy on hard entities, expressive controllability, and general TTS quality — and uses three benchmarks accordingly. (i) **DRUG_REF** is our pharmaceutical-entity pronunciation benchmark of 7,500 clips, generated from the catalog in Section 4.2 as 300 drugs $\times$ 5 templates $\times$ 5 generations. Its held-out DRUG-OOD split contains the 30 drug names excluded from DPO preference construction. (ii) **EXP_EMO** is a 100-clip benchmark for explicit emotion-tag controllability, described in detail below. (iii) **Seed-TTS** (Anastassiou et al., 2024) is the public reference we use for general text-conditioned synthesis and voice cloning. Together, the three benchmarks let us audit each design choice against a metric outside the one it is optimized for.

**Drug-pronunciation evaluation.** We use a multimodal judge — Gemini-3.1-Pro (The Gemini Team, 2026) at $T = 0$ — to score each generated clip against a human reference recording of the correct pronunciation. The judge returns one of three labels (*correct* / *partial* / *wrong*) along with a short acoustic justification, and we report the strict *correct* rate as mean $\pm$ 5-run $\sigma$ over the five-generation partition. Unless explicitly noted, drug-pronunciation rows use inference-time reference biasing (Section 3.4); the same protocol is applied to DRUG-OOD and to the cross-system rows in Table 1. As a recognizer-side complement to the judge, we also report entity-span phoneme error rate (PER) using NeMo Forced Aligner and the wav2vec 2.0 phoneme recognizer.

**Expressivity evaluation (EXP_EMO).** EXP_EMO is a 100-clip clinical-conversational test set covering five emotion classes (*neutral, happy, sad, angry, disgust*) with 20 sentences per class, each prefixed by the matching `<|emotion_*|>` token. We score every clip on two signals: (i) an audio-only blind LLM judge (Gemini-3.1-Pro at $T = 0$, given no transcript or target tag, calibrated on a 20-clip pilot to $\geq 85\,\%$ all-confidence accuracy) that predicts the expressed emotion; and (ii) DeEAR arousal Cohen's $d$ (Lin et al., 2025), reported as mean $\bar{d}$ against the neutral baseline along with mean $|\Delta F_0|$ in semitones (Table 5).

**Other metrics.** To detect regressions outside the pronunciation and expressivity objectives, we additionally report Allosaurus speaking rate in phones/s (Li et al., 2020) and Seed-TTS-Eval WER and speaker similarity (SIM) (Anastassiou et al., 2024) on Seed-TTS Test-EN. We also report two complementary predicted-naturalness scores: UTMOSv2 (Baba et al., 2024) on the drug-pronunciation set, where the text content is fixed and only the entity-span realization varies, and DNSMOS P.808 (Reddy et al., 2021) on EXP_EMO, which exercises broader prosody and emotion-tagged delivery.

| Model | Drug-pron ↑ | PER ↓ | UTMOSv2 ↑ | Rate | WER ↓ | SIM ↑ |
|---|---|---|---|---|---|---|
| phon_0 | 4.2 ± 0.8 | 74.81 | 3.43 | 10.22 | 2.01 | 67.1 |
| phon_20 | 86.7 ± 0.7 | 22.39 | 3.48 | 11.18 | 1.81 | 67.2 |
| phon_40 | 90.9 ± 1.0 | 21.72 | 3.48 | 11.07 | 1.87 | 67.1 |
| phon_60 | 90.7 ± 0.6 | 22.20 | 3.49 | 11.15 | **1.80** | 67.2 |
| phon_80 | 91.5 ± **0.3** | 21.41 | 3.49 | 11.12 | 1.92 | 66.6 |
| phon_100 | 91.8 ± 0.4 | 21.13 | 3.50 | 11.04 | 1.81 | 65.6 |
| phon_60 + DPO only | 89.7 ± 0.8 | **12.84** | 3.49 | 10.50 | 1.86 | 67.6 |
| phon_60 + DPO + SFT | **93.1** ± 0.4 | 14.61 | 3.54 | 10.93 | 1.83 | 67.2 |
| *closed-source TTS baseline* | | | | | | |
| ElevenLabs Turbo v2 | 73.4 ± 0.5 | 20.51 | 3.46 | 12.00 | — | — |
| Inworld TTS-2 | 51.5 ± 1.1 | 26.06 | 3.69 | 9.83 | — | — |
| *open-source TTS baselines* | | | | | | |
| FireRedTTS-2 | 10.6 ± 1.0 | 78.82 | 3.05 | 8.20 | 1.95 | 66.5 |
| CosyVoice3 | 82.6 ± 1.2 | 25.52 | **3.62** | 10.61 | 2.22 | 72.0 |
| VoxCPM2 | 86.1 ± 0.7 | 22.03 | 3.45 | 11.68 | 1.84 | **75.3** |

Table 1: Pronunciation, acoustic quality, and Seed-TTS-Eval results on the 7,500-clip drug-pronunciation benchmark (300 drugs × 5 sentences × 5 generations). *Drug-pron* is Gemini-judged drug-name correctness (%) reported as mean ± 5-run $\sigma$; PER is Wav2Vec2 entity-span phoneme error rate (%); *Rate* is Allosaurus phones/s. WER and SIM are Seed-TTS-Eval Test-EN metrics. All entries use inference-time reference biasing (Section 3.4). Baseline WER/SIM values for open-source systems are reported from Zhou et al. (2025); closed-source systems are not part of Seed-TTS-Eval, so those cells are dashed.

## 5 Results and Discussion

### 5.1 SFT Unlocks Phoneme Control

The post-trained FireRedTTS-2 checkpoint has little ability to interpret ARPAbet input directly: the PHON_0 control, trained without phoneme augmentation, reaches only 4.2 % *Drug-pron* accuracy and 74.81 % entity PER. Adding even a small amount of phoneme-augmented SFT data changes this behavior sharply. At 10 % phoneme annotation, *Drug-pron* accuracy rises to 89.5 % and PER falls to 22.59 %.

Increasing the phoneme annotation rate beyond this point gives a plateau rather than a monotonic improvement. The PHON_40–PHON_100 checkpoints all cluster near 91 % *Drug-pron* accuracy and 21–22 % PER. We therefore treat the sweep as a model-selection tool rather than a quantity to maximize blindly; the operating-point selection is justified after the biasing ablation below.

### 5.2 Reference Biasing Stabilizes Known Entities

We evaluate the inference-time biasing of Section 3.4. Table 2 shows it improves every non-trivial phoneme-adapted checkpoint by 6–9 points and tightens 5-run $\sigma$ by 3–10×. All drug-pronunciation rows in Table 1 use biasing. The setting assumes a reference pronunciation is available, so biasing should be read as complementary to and not a replacement for phoneme adaptation.

| Variant | w/ biasing ↑ | no biasing ↑ |
|---|---|---|
| phon_0 | 4.2 ± 0.8 | 2.7 ± 0.1 |
| phon_60 | 90.7 ± 0.6 | 84.6 ± 3.4 |
| phon_100 | 91.8 ± 0.4 | 83.2 ± 3.4 |

Table 2: Inference-time biasing ablation on representative SFT checkpoints (Drug-pron %, mean ± 5-run $\sigma$).

We select **PHON_60** on a composite ranking that weights pronunciation, expressivity, and general synthesis quality: it has the best Seed-TTS-Eval WER (1.80 %), ties for best blind-emotion accuracy (90.0 %), is within 1.1 pp of the highest drug-pron rate at tight 5-run $\sigma$, and has the second-highest DeEAR $\bar{d}$ (2.19). We initialized the DPO from this checkpoint.

### 5.3 Effect of Loss Mask on DPO

We ablate the entity-span-centered loss mask against a vanilla autoregressive mask during DPO (Table 3). The entity-centered mask improves PER by 43% relative, at the cost of lowering Allosaurus speaking rate from 11.01 to 10.52. Henceforth, all DPO results use the entity-centered mask.

### 5.4 DPO Improves Pronunciation but Changes Delivery

DPO provides the final pronunciation improvement on top of PHON_60. The DPO+SFT-replay

| Model | PER ↓ | Phones/sec |
|---|---|---|
| SFT (PHON_60) | 20.96 % | 11.15 |
| DPO (vanilla mask) | 18.43 % | 11.01 |
| DPO (entity-centered mask) | **10.37 %** | 10.52 |

Table 3: Effect of DPO loss mask on DRUG-OOD.

| Model | Drug-pron ↑ | PER ↓ | Emo ↑ | Rate | $\bar{d}$ |
|---|---|---|---|---|---|
| phon_60 | 87.25 | 20.96 | 90.0 | 11.15 | 2.19 |
| DPO only | 95.71 | **10.37** | 81.0 | 10.50 | 1.30 |
| DPO+replay | **96.19** | 11.53 | 79.0 | 10.93 | 1.48 |

Table 4: Generalization on DRUG-OOD (30 held-out drug names, 1,890 clips). Columns follow Table 1: *Drug-pron* is Gemini-judged strict correct rate (%); PER is Wav2Vec2 entity-span phoneme-error rate (%); *Emo*: blind LLM-judge emotion accuracy (%); $\bar{d}$: DeEAR effect size; *Rate*: phones/s.

| Model | Blind | $\bar{d}$ | $\lvert\Delta F_0\rvert$ | DNSMOS |
|---|---|---|---|---|
| phon_60 | 90.0 | 2.19 | 0.70 | 4.01 |
| phon_60 + DPO | 81.0 | 1.30 | 0.72 | **4.09** |
| phon_60 + DPO+replay | 79.0 | 1.48 | 0.33 | 4.02 |
| *closed-source* | | | | |
| Inworld TTS-2 (tags) | 89.0 | 2.53 | 0.33 | **4.18** |
| ElevenLabs Turbo v2 | 89.0 | 1.78 | 0.49 | 4.09 |
| ElevenLabs v3 (tags) | **98.0** | 3.51 | 1.57 | 3.98 |
| *open-source* | | | | |
| FireRedTTS-2 base | 81.0 | 1.25 | 0.30 | 3.99 |
| CosyVoice3 | 96.0 | **3.96** | 1.15 | 3.93 |
| VoxCPM2 | 97.0 | 1.83 | **2.20** | 3.78 |

Table 5: Expressivity results on the 100-utterance emotion-tag test set. Blind emo is audio-only LLM-judge target-match rate. $\bar{d}$ is mean DeEAR (Lin et al., 2025) arousal Cohen's $d$ against neutral. $\overline{|\Delta F_0|}$ is mean absolute pitch shift in semitones.

checkpoint reaches 93.1 % Gemini accuracy (vs. 90.7 % for SFT alone) and reduces entity PER from 22.20 % to 14.61 % — the highest drug-pronunciation score in our evaluation and 7.0 points above VoxCPM2, the strongest open-source baseline. The DPO-only variant achieves the lowest full-benchmark PER (12.84 %) but lower judge-facing accuracy, showing that optimizing recognizer-facing PER alone is not identical to maximizing judge-facing correctness.

The direct DPO generalization test is DRUG-OOD, the 30 names held out from preference construction. Table 4 gives the full multi-objective view: DPO+SFT-replay lifts Gemini accuracy from 87.25 % to 96.19 % (+8.94 pp) and roughly halves entity PER (20.96 → 11.53 %), showing that DPO generalizes beyond the 270 training drugs rather than memorizing them. Replay-SFT has slightly higher OOD PER than DPO-only but is the preferred operating point: it has the highest Gemini-judged drug-pronunciation rate on both the in-distribution benchmark (93.1 % vs. 89.7 %) and the OOD split (96.2 % vs. 95.7 %), and it stays closer to the SFT baseline on speaking rate and DeEAR arousal.

The same comparison exposes a side effect: DPO slows generation (Allosaurus rate 11.15 → 10.93/10.50 phones/s for replay/DPO-only) and reduces blind-emotion accuracy and $\bar{d}$ on the OOD set, mirroring the in-distribution expressivity regression in Section 5.6. Two non-exclusive mechanisms are plausible: slower speech spreads phonetic content across more codec frames (easing both model and recognizer), and wav2vec 2.0 confidence rises on temporally dilated entity spans (biasing pair selection toward slower candidates). The pattern motivates measuring rate and prosody alongside PER when using recognizer-derived preferences.

### 5.5 General TTS Quality Is Preserved

Phoneme adaptation does not degrade general text-conditioned synthesis on the automatic metrics we report. Across the phoneme-rate sweep, Seed-TTS-Eval WER stays in a tight 1.80–2.01 % band; the DPO+SFT-replay model matches PHON_60 on both WER (1.83 % vs. 1.80 %) and speaker similarity (both 67.2); and UTMOSv2 stays flat across SFT and DPO. We make this claim at the resolution of recognition, speaker-identity, and naturalness predictors — not at the level of a human listening study.

### 5.6 Expressivity Reveals the DPO Trade-off

Table 5 places our PHON_60 stack alongside six open- and closed-source TTS systems on the EXP_EMO benchmark. The question we want to answer is whether phoneme-aware SFT and pronunciation-focused DPO preserve emotion-conditioned synthesis when prompts carry an explicit `<|emotion_*|>` tag. PHON_60 SFT alone reaches 90 % blind-emotion accuracy and a DeEAR (Lin et al., 2025) effect size of $\bar{d} = 2.19$, on par with the closed-source low-latency systems — phoneme adaptation by itself does not erase expressivity.

Three patterns stand out in Table 5. (i) *Blind LLM-judge accuracy saturates into two bands.* The low-latency real-time systems — Inworld,

ElevenLabs Turbo v2, and PHON_60 — cluster at 89–90 % blind accuracy, while the higher-latency expressivity-focused systems — ElevenLabs v3, CosyVoice3, VoxCPM2 — form an upper band at 96–98 %. (ii) *DeEAR and* $|\Delta F_0|$ *discriminate more sharply.* CosyVoice3 leads arousal ($\bar{d} = 3.96$, above ElevenLabs v3 at 3.51) and VoxCPM2 leads pitch shift (2.20 st, above v3 at 1.57); on both axes PHON_60 ($\bar{d} = 2.19$) sits with the low-latency systems rather than with the expressivity-focused band. (iii) *DNSMOS is tightly clustered* (3.78–4.18) across every row, so phoneme adaptation and DPO preserve naturalness at the level of closed-source flagships. The FireRedTTS-2 base row — our backbone without the expressive-augmentation SFT — isolates that contribution: it drops to 81 % blind accuracy and $\bar{d} = 1.25$, showing that the tag-mapping training in our SFT mixture is what unlocks the emotion-tag interface.

DPO, however, erodes expressive control. PHON_60+DPO drops blind-emotion accuracy from 90 % to 79 %, $\bar{d}$ from 2.19 to 1.48, and $|\Delta F_0|$ from 0.70 st to 0.33 st. Since Seed-TTS-Eval intelligibility, speaker similarity, and UTMOSv2 are unchanged by DPO, this regression is specific to expressive delivery rather than a general quality collapse — the pronunciation gain comes at a measurable expressivity cost.

## 6 Limitations

The recipe discussed in this work is English-only and ARPAbet-only: we have not tested cross-lingual transfer or other phoneme alphabets, and dense phoneme prompting may still introduce naturalness or prosody regressions outside our test domain. We also have not tested transfer to TTS architectures other than FireRedTTS-2, and the recipe is sensitive to phoneme annotation quality during SFT.

A second class of risk is metric-induced bias during DPO. Because chosen and rejected samples are scored with an automatic phoneme recognizer, the policy can drift toward features the recognizer rewards rather than features human listeners do. We see one such effect on speaking rate: the DPO-adapted model is slower than the SFT-only checkpoint even as recognizer-side PER improves. A plausible explanation is that slower speech spreads each phone across more codec frames, which can ease both TTS generation and recognizer scoring; we treat this as a hypothesis rather than a confirmed cause. The held-out DRUG-OOD split rules out drug-identity memorization, but it does not rule out this class of recognizer-side overfitting.

## 7 Conclusion

In this work we presented a practical recipe for retrofitting explicit phoneme controllability onto a pretrained LLM-based TTS model. The final system stacks phoneme-aware SFT, pronunciation-focused DPO with SFT replay, and inference-time reference biasing, reaches 93.1 % strict correctness on the drug-pronunciation benchmark, and roughly halves held-out DRUG-OOD entity PER from 20.96 % to 11.53 %. On the same benchmark it beats every open- and closed-source TTS system we evaluated.

These gains are not free. Seed-TTS-Eval WER and SIM, UTMOSv2, and DNSMOS all stay stable, but DPO shifts emotional delivery: blind-emotion accuracy drops from 90 % to 79 % and DeEAR $\bar{d}$ from 2.19 to 1.48. We therefore argue for evaluating phoneme adaptation as a multi-objective problem — reporting pronunciation accuracy alongside general quality, speaker similarity, speaking rate, and expressivity — rather than treating any single metric as the headline. Two natural next steps are curriculum-style preference signals that recover expressivity without sacrificing pronunciation, and a multi-system human listening study to validate the automatic-judge findings.

## 8 Ethical Considerations

Sharpening phoneme reading in a voice-cloning TTS model also makes synthetic speech more convincing for impersonation, harassment, or harmful-content generation. Deployment should require speaker consent, access controls, abuse monitoring, and provenance mechanisms such as watermarking or synthetic-speech disclosure. In high-stakes domains, correct pronunciation should not be confused with factual or clinical correctness: generated content still needs domain-specific validation. Our automatic judges may also not reflect listener judgments across accents, dialects, or languages.

## Acknowledgements

We thank the Inworld TTS team for providing free API credits that enabled the evaluation of their Realtime TTS-2 model in this paper.

## References


Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, Mingqing Gong, Peisong Huang, Qingqing Huang, Zhiying Huang, Yuanyuan Huo, Dongya Jia, Chumin Li, Feiya Li, Hui Li, and 27 others. 2024. Seed-tts: A family of high-quality versatile speech generation models. *Preprint*, arXiv:2406.02430.

Kaito Baba, Wataru Nakata, Yuki Saito, and Hiroshi Saruwatari. 2024. The t05 system for the voicemos challenge 2024: Transfer learning from deep image classifier to naturalness mos prediction of high-quality synthetic speech. *Preprint*, arXiv:2409.09305.

Carnegie Mellon University Speech Group. The CMU Pronouncing Dictionary. http://www.speech.cs.cmu.edu/cgi-bin/cmudict?in=C+M+U+Dictionary. Accessed: 2026-05-24.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. 2024. XTTS: a massively multilingual zero-shot text-to-speech model. *arXiv preprint arXiv:2406.04904*.

Ziqi Chen, Gongyu Chen, Yihua Wang, Chaofan Ding, Zihao Chen, and Wei-Qiang Zhang. 2025. DiaMoE-TTS: A unified IPA-based dialect TTS framework with mixture-of-experts and parameter-efficient zero-shot adaptation. *arXiv preprint arXiv:2509.22727*.

Christopher Cieri, David Miller, and Kevin Walker. 2004. Fisher english training speech part 1. Linguistic Data Consortium, LDC2004S13.

Christopher Cieri, David Miller, and Kevin Walker. 2005. Fisher english training speech part 2. Linguistic Data Consortium, LDC2005S13.

Giulia Comini, Sam Ribeiro, Fan Yang, Heereen Shim, and Jaime Lorenzo-Trueba. 2023. Multilingual context-based pronunciation learning for text-to-speech. In *Interspeech 2023*, pages 631–635.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, Keyu An, Guanrou Yang, Yabin Li, Yanni Chen, Zhifu Gao, Qian Chen, Yue Gu, Mengzhe Chen, Yafeng Chen, and 3 others. 2025. CosyVoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*.

Jason Fong, Daniel Lyth, Gustav Eje Henter, Hao Tang, and Simon King. 2022. Speech audio corrector: using speech from non-target speakers for one-off correction of mispronunciations in grapheme-input text-to-speech. In *Interspeech 2022*, pages 1213–1217.

Jason Fong, Hao Tang, and Simon King. 2023. Spell4tts: Acoustically-informed spellings for improving text-to-speech pronunciations. In *12th ISCA Speech Synthesis Workshop (SSW2023)*, pages 8–13.

Jason Fong, Jason Taylor, and Simon King. 2020. Testing the limits of representation mixing for pronunciation correction in end-to-end speech synthesis. In *Interspeech 2020*, pages 4019–4023.

Mutian He, Jingzhou Yang, Lei He, and Frank Soong. 2022. Neural lexicon reader: Reduce pronunciation errors in end-to-end TTS by leveraging external textual knowledge. In *Interspeech 2022*, pages 441–445.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, Xinyu Zhang, Pei Zhang, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. 2026. Qwen3-TTS technical report. *arXiv preprint arXiv:2601.15621*.

Ye Jia, Heiga Zen, Jonathan Shen, Yu Zhang, and Yonghui Wu. 2021. PnG BERT: Augmented BERT on phonemes and graphemes for neural TTS. In *Interspeech 2021*, pages 151–155.

Daniel Jurafsky and James H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 1st edition. Prentice Hall. ARPAbet notation system, pp. 94–95.

Kyle Kastner, João Felipe Santos, Yoshua Bengio, and Aaron C. Courville. 2019. Representation mixing for TTS synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Adrian Łańcucki. 2021. FastPitch: Parallel text-to-speech with pitch prediction. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Ryan Langman, Xuesong Yang, Paarth Neekhara, Shehzeen Hussain, Edresson Casanova, Evelina Bakhturina, and Jason Li. 2025. HiFiTTS-2: A large-scale high bandwidth speech dataset. *arXiv preprint arXiv:2506.04152*.

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R. Mortensen, Graham Neubig, Alan W. Black, and Florian Metze. 2020. Universal phone recognition with a multilingual allophone system. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8249–8253.

Yinghao Aaron Li, Cong Han, Vinay S. Raghavan, Gavin Mischler, and Nima Mesgarani. 2023. StyleTTS 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. *Advances in Neural Information Processing Systems*.

Zhiyu Lin, Jingwen Yang, Jiale Zhao, Meng Liu, Sunzhu Li, and Benyou Wang. 2025. Decoding the ear: A framework for objectifying expressiveness from human preference through efficient alignment. *arXiv preprint arXiv:2510.20513*.

Linh The Nguyen, Thinh Pham, and Dat Quoc Nguyen. 2023a. XPhoneBERT: A pre-trained multilingual model for phoneme representations for text-to-speech. *arXiv preprint arXiv:2305.19709*.

Tu Anh Nguyen, Wei-Ning Hsu, Antony D'Avirro, Bowen Shi, Itai Gat, Maryam Fazel-Zarani, Tal Remez, Jade Copet, Gabriel Synnaeve, Michael Hassid, Felix Kreuk, Yossi Adi, and Emmanuel Dupoux. 2023b. EXPRESSO: A benchmark and analysis of discrete expressive speech resynthesis. *arXiv preprint arXiv:2308.05725*.

NVIDIA. 2025. Nemo forced aligner (NFA). NVIDIA NeMo Framework User Guide.

Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O. Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. 2018. Deep voice 3: Scaling text-to-speech with convolutional sequence learning. In *International Conference on Learning Representations (ICLR)*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*.

Chandan K. A. Reddy, Vishak Gopal, and Ross Cutler. 2021. DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP 2021 – 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6493–6497. IEEE.

Yi Ren and Danica J. Sutherland. 2024. Learning dynamics of LLM finetuning. *arXiv preprint arXiv:2407.10490*.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. 2018. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

The Gemini Team. 2026. Gemini 3.1 pro: A smarter model for your most complex tasks. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/. Google Keyword Blog. Accessed: 2026-05-24.

Jinchuan Tian, Chunlei Zhang, Jiatong Shi, Hao Zhang, Jianwei Yu, Shinji Watanabe, and Dong Yu. 2024. Preference alignment improves language model-based TTS. *arXiv preprint arXiv:2409.12403*.

Wenming Tu, Guanrou Yang, Ruiqi Yan, Wenxi Chen, Ziyang Ma, Yipeng Kang, Kai Yu, Xie Chen, and Zilong Zheng. 2025. UltraVoice: Scaling fine-grained style-controlled speech conversations for spoken dialogue models. *arXiv preprint arXiv:2510.22588*.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. 2017. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.

Kun Xie, Feiyu Shen, Junjie Li, Fenglong Xie, Xu Tang, and Yao Hu. 2025. Fireredtts-2: Towards long conversational speech generation for podcast and chatbot. *arXiv preprint arXiv:2509.02020*.

Qiantong Xu, Alexei Baevski, and Michael Auli. 2021. Simple and effective zero-shot cross-lingual phoneme recognition. *arXiv preprint arXiv:2109.11680*.

Guangyan Zhang, Kaitao Song, Xu Tan, Daxin Tan, Yuzi Yan, Yanqing Liu, Gang Wang, Wei Zhou, Tao Qin, Tan Lee, and Sheng Zhao. 2022. Mixed-phoneme BERT: Improving BERT with mixed phoneme and sup-phoneme representations for text to speech. In *Interspeech 2022*, pages 456–460.

Leying Zhang, Yao Qian, Xiaofei Wang, Manthan Thakker, Dongmei Wang, Jianwei Yu, Haibin Wu, Yuxuan Hu, Jinyu Li, Yanmin Qian, and Sheng Zhao. 2025a. CoVoMix2: Advancing zero-shot dialogue generation with fully non-autoregressive flow matching. In *Advances in Neural Information Processing Systems*.

Leying Zhang, Yao Qian, Long Zhou, Shujie Liu, Dongmei Wang, Xiaofei Wang, Midia Yousefi, Yanmin Qian, Jinyu Li, Lei He, Sheng Zhao, and Michael Zeng. 2024. CoVoMix: Advancing zero-shot speech generation for human-like multi-talker conversations. In *Advances in Neural Information Processing Systems*.

Xueyao Zhang, Yuancheng Wang, Chaoren Wang, Ziniu Li, Zhuo Chen, and Zhizheng Wu. 2025b. Advancing zero-shot text-to-speech intelligibility across diverse domains via preference alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12251–12270.

Yuqian Zhang, Donghua Yu, Zhengyuan Lin, Botian Jiang, Mingshu Chen, Yaozhou Jiang, Yiwei Zhao, Yiyang Zhang, Yucheng Yuan, Hanfu Chen, Kexin Huang, Jun Zhan, Cheng Chang, Zhaoye Fei, Shimin Li, Xiaogui Yang, Qinyuan Cheng, and Xipeng Qiu. 2026. MOSS-TTSD: Text to spoken dialogue generation. *arXiv preprint arXiv:2603.19739*.

Yixuan Zhou, Guoyang Zeng, Xin Liu, Xiang Li, Renjie Yu, Ziyang Wang, Runchuan Ye, Weiyue Sun, Jiancheng Gui, Kehan Li, Zhiyong Wu, and Zhiyuan Liu. 2025. VoxCPM: Tokenizer-free TTS for context-aware speech generation and true-to-life voice cloning. *arXiv preprint arXiv:2509.24650*.

# A Gemini-as-Judge Prompts

## A.1 Drug-Pronunciation Judge

The drug-pronunciation eval (Table 1) calls Gemini-3.1-Pro at $T = 0$ with two audio inputs (AUDIO 1 = reference recording of the drug name spoken in isolation; AUDIO 2 = the TTS-generated clip) and the prompt below. The IPA block is shown only as descriptive context for error annotations and is explicitly excluded from the scoring decision.

```
You are evaluating whether a TTS system
correctly pronounced a medical/drug term.
Focus ONLY on the pronunciation of the
target term. Ignore the surrounding
sentence quality, speaker voice, pacing,
and expressiveness.

Target term: {drug_name}
Reference IPA (for error description only
  -- do NOT use this to determine your
  score): {ipa}
AUDIO 1: Reference recording of
  "{drug_name}" spoken in isolation.
AUDIO 2: TTS synthesis of the sentence:
  "{sentence}"

## Scoring instructions

Base your score SOLELY on how the
pronunciation in AUDIO 2 compares to
AUDIO 1. Do NOT score against the IPA
text -- the reference audio is the only
ground truth.

Rate the pronunciation of "{drug_name}"
in AUDIO 2:
  3 = Correct: matches the reference
      audio clearly and correctly
  2 = Flawed but recognizable: identifiable
      as the same drug, but has audible
      errors vs the reference
  1 = Wrong or unintelligible: sounds like
      a different word, is garbled, or is
      skipped entirely

Output structured JSON:
  - If 3: {"score": 3}
  - If 2/1: include error_type,
    heard_as (phonetic description), and
    confidence (high|medium|low).
```

The corpus-level strict correct rate (% of clips with `score=3`) is reported in Table 1 as mean $\pm$ 5-run $\sigma$ from a 5-generation partition. The same prompt is applied verbatim to the DRUG-OOD split and to every cross-system row.

## A.2 Blind Emotion Judge

The blind emotion accuracy reported in Table 5 is produced by Gemini-3.1-Pro at $T = 0$ with a single audio input and the following prompt. The judge sees no transcript, no target tag, and no metadata of any kind. The prompt was iterated on a 20-clip calibration set until $\approx 85\,\%$ all-confidence accuracy and elimination of obvious failure modes (lexical inference, hallucinated acoustic events).

```
You are an expert acoustic analyst
evaluating a Text-to-Speech audio clip.

TASK: classify the speaker's emotional
DELIVERY based on VOCAL CHARACTERISTICS.

You will hear a single short audio clip.
You will NOT see a transcript and will
NOT be told what the speaker is saying.
Judge entirely from acoustic cues:
  * Pitch level (high / mid / low)
  * Pitch variation (flat / moderate /
    wide range)
  * Tempo (slow / normal / fast)
  * Voice quality (breathy / clear /
    harsh / tense / relaxed)
  * Energy / loudness (soft / normal /
    forceful)

CATEGORIES (pick exactly one):
  - neutral: Conversational, flat-to-
    moderate prosody, no strong emotional
    coloring. THIS IS THE DEFAULT --
    pick this when in doubt.
  - happy: Upbeat, raised pitch or wider
    pitch range, lighter voice quality,
    often faster tempo, smiling tone.
  - sad: Lowered pitch register, slower
    tempo, often quieter or breathier,
    falling intonation.
  - angry: Forceful energy, sharper
    consonants, harsh or tense voice
    quality, often louder.
  - disgust: Pulled-back or constricted
    voice quality, often falling
    intonation, nasal or guttural
    undertone.

CRITICAL RULES:
  1. Default to "neutral." Only assign a
     non-neutral category when there are
     CLEAR, IDENTIFIABLE acoustic cues
     that drive the assignment.
  2. Confidence calibration:
     - "high"   = you would bet money on
       this. Multiple cues align.
     - "medium" = leading interpretation
       but you could see "neutral."
     - "low"    = guess based on weak
       cues. Prefer "neutral" instead.

Output structured JSON matching the
response schema. Be specific in
acoustic_features -- generic descriptions
like "natural-sounding" are not useful.
```

The prompt was iterated on a 20-clip calibration set drawn from a single phon_X checkpoint until the judge reached $\geq 85\,\%$ all-confidence target-match on a held-out pilot and obvious failure modes (lexical-context shortcuts, hallucinated paralinguistic events) were eliminated. Predictions are post-hoc compared to the intended target tag.

## B Per-Emotion DeEAR Breakdown

Table 6 reports per-emotion DeEAR arousal Cohen's $d$ vs. neutral for the phon_60 SFT baseline and the phon_60 + DPO + SFT-replay checkpoint on the same EXP_EMO set ($n = 20$ clips per emotion). DPO compresses three of the four non-neutral cells (*angry*, *happy*, *disgust* all drop substantially), while *sad* actually rises — consistent with the preference-data narrative in Section 5 (the entity-PER objective favors calm, articulated deliveries, which are easier to produce in the *sad* register than in the high-arousal *angry* register).

| Emotion | phon_60 | + DPO+replay | Δ |
|---|---|---|---|
| happy | 2.31 | 1.15 | −1.16 |
| sad | 1.34 | 2.13 | +0.79 |
| angry | **2.62** | 1.15 | −1.47 |
| disgust | 2.48 | 1.50 | −0.98 |

Table 6: DeEAR arousal Cohen's $d$ vs. neutral, per emotion ($n = 20$ per cell). DPO reduces three of four cells; the largest drop is on *angry* ($-1.47$).

## C Per-Drug Accuracy Distribution (OOD-30)

Because the DPO rollout pool draws from the same 270-drug training partition used for SFT, reporting per-drug rates on the full 300-drug benchmark would conflate memorization with generalization. Table 7 therefore restricts the breakdown to the 30-drug held-out OOD set, with 63 generations per drug ($\geq 5$ templates $\times$ multiple seeds). Three patterns are visible. (i) *SFT alone leaves a long tail*: one drug below $40\,\%$ (*Beyfortus*, $33\,\%$) and 6 more in $[60, 80)$. (ii) *DPO eliminates catastrophic failures*: the worst OOD drug for DPO-only is *sacubitril* at $65\,\%$, and the entire tail below $60\,\%$ disappears. (iii) *DPO + SFT-replay pushes every OOD drug to* $\geq 80\,\%$: $20/30$ drugs sit in $[80, 100)$ and $10/30$ are perfect. The corpus rates ($87.2\,\% \rightarrow 95.7\,\% \rightarrow 96.2\,\%$) align with the named-cell numbers in Table 4.

## D Audio Samples

We provide representative generated speech samples for our adapted models and the evaluated open- and closed-source TTS baselines on an anonymous demo page: https://phoneme-control.github.io/. The samples are intended to complement the automatic pronunciation, quality, and expressivity metrics reported in the main paper by allowing listeners to compare pronunciation behavior across systems directly.

| Per-drug rate | SFT | +DPO | +DPO+replay |
|---|---|---|---|
| [0, 20) | 0 | 0 | 0 |
| [20, 40) | 1 | 0 | 0 |
| [40, 60) | 0 | 0 | 0 |
| [60, 80) | 6 | 2 | 0 |
| [80, 100) | 19 | 16 | 20 |
| 100 (perfect) | 4 | **12** | 10 |
| total drugs | 30 | 30 | 30 |
| corpus rate (%) | 87.2 | 95.7 | **96.2** |

Table 7: Per-drug strict correct-rate distribution on the OOD-30 held-out set (63 generations per drug, 1,890 clips total per system). SFT = phon_60; +DPO = phon_60+DPO only; +DPO+replay = phon_60+DPO+SFT-replay. DPO removes the $< 60\,\%$ tail; replay then lifts the remaining $[60, 80)$ drugs to $\geq 80\,\%$.