# LYRICO: CONTROLLABLE CHINESE LYRIC GENERATION WITH LARGE LANGUAGE MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Controllable Chinese lyric generation requires models to satisfy both content requirements and fine-grained structural constraints, including section layout, line count, line length, word-group lengths, and rhyme. We present *LyriCo*, a structured-text framework that encodes structural specifications as control tokens and fine-tunes decoder-only language models on annotated lyrics. Its data pipeline combines classifier-based genre annotation, theme labels extracted by a large language model, text-to-speech (TTS)-inspired Chinese word segmentation, and section annotations derived from music structure analysis and lyric alignment. We introduce two lightweight mechanisms to strengthen structural control: increased loss weights on control tokens and a line-local attention bias toward relevant control tokens. On a benchmark of 500 briefs spanning Pop and gufeng, ablation studies across seven configurations at the 1.5B parameter scale evaluate the contributions of both mechanisms. Compared with the standard supervised fine-tuning baseline, the combined configuration improves line-count accuracy from 93.40% to 96.80%, section accuracy from 95.19% to 98.80%, and word-format accuracy from 97.96% to 98.52%. With a 14B backbone, LyriCo achieves higher line-length, rhyme, and word-format accuracy than prompted GPT-5.5 and DeepSeek-V4-Pro on the same benchmark, while tying with GPT-5.5 for the highest line-count accuracy. These results show that emphasizing explicit control signals can improve compliance with structural constraints in Chinese lyric generation.

**One theme, two lyric styles**

The anxious, wavering state of mind in love
– caught between hope and fear, unable to decide.

Verse: 4 lines / Chorus: 6 lines
Rhyme: i / ü (class 3), all lines

| POP 《该不该》<br>Should I? | SHARED TARGET<br>Chars | <br>Groups | GUFENG 《难生恨》<br>Hard to Hate |
|---|---|---|---|
| **VERSE** | | | |
| 窗外\|雨在滴<br>Outside, the rain keeps dripping | 5 | 2+3 | 那日\|听莺啼<br>That day, I heard the orioles' song |
| 明明\|想见你<br>Clearly, I want to see you | 5 | 2+3 | 今朝\|闻飞絮<br>This morning, catkins fill the air |
| **CHORUS** | | | |
| 到底该不该\|真心掏给你<br>Should I hand you my bare heart? | 10 | 5+5 | 我无悔相遇\|相知到相惜<br>No regret meeting, knowing, treasuring you |
| 互相拉扯\|实在压抑\|让人窒息<br>Push-pull, so stifling, it suffocates | 12 | 4+4+4 | 纵使千般\|思绪万缕\|痛彻心底<br>A thousand thoughts, piercing deep |

Figure 1: Chinese lyric generation in Pop and *gufeng* under shared content requirements and structural constraints. The central columns specify line lengths and word-group lengths; bars separate word groups, and colored final characters mark rhyme. Verse and chorus labels indicate section layout. English translations are provided for reference. Complete ten-line examples appear in Appendix C.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

# 1 INTRODUCTION

Controllable Chinese lyric generation coordinates content requirements with structural constraints at several scales. Numerical planning remains challenging (Sun et al., 2023), and compositional benchmarks distinguish constraints at different textual scales (Yao et al., 2024).

A content brief specifies theme and genre, with structural constraints on section layout, line count, line length, word-group lengths, and rhyme. *Word format* denotes a sequence of character-group lengths, not necessarily lexical words. Line length is measured by Chinese-character count, our operational proxy for syllable count.

Explicit structural representations have substantial precedents. SongNet (Li et al., 2020) and ChipSong (Liu et al., 2022) expose format and rhyme information through symbols and adapted attention. More recently, Chae et al. (2025) combine full-song section control with syllable counts at multiple granularities. Our question is therefore how a decoder-only large language model (LLM) uses jointly specified Chinese lyric constraints, including genre, rhyme, and internal character groups. In particular, how much control comes from emphasizing control tokens after supervised fine-tuning has already taught the output format?

*LyriCo* is a structured-text framework that interleaves lyric text with control tokens encoding structural specifications. Its pipeline combines classifier-based genre annotation, LLM-extracted theme labels, TTS-inspired Chinese word segmentation, and section annotations from music structure analysis and lyric alignment. On this representation, **control-token loss weighting** emphasizes structural prediction errors, and **line-local attention bias** favors relevant control tokens within each line during training and generation.

We evaluate 500 briefs spanning Pop and *gufeng*. Seven 1.5B configurations compare standard supervised fine-tuning (SFT) with the two mechanisms. The combination improves line-count, section, and word-format accuracy over standard SFT. At 14B, LyriCo exceeds prompted GPT-5.5 and DeepSeek-V4-Pro on line-length, rhyme, and word-format accuracy, and ties with GPT-5.5 on line-count accuracy. The complete recipe also transfers to GLM-9B. Separately, LyriCo-14B attains 99.93% word-format accuracy on 506 cases from ChipSong's released resources.

**Contributions.** Figure 1 illustrates Pop and *gufeng* examples under a shared structural brief.

- A structured-text framework and annotation pipeline for Chinese lyric generation with content conditioning and multiple explicit structural controls.
- Two lightweight mechanisms: control-token loss weighting and line-local attention bias, evaluated against standard supervised fine-tuning in seven 1.5B configurations.
- Evaluation on 500 briefs across two backbone families, 506 cases from ChipSong's released resources, and complementary annotation and content-adherence checks.

# 2 RELATED WORK

## 2.1 LYRIC AND POETRY GENERATION

Lyric and poetry generation has a long history of explicit structural control. SongNet (Li et al., 2020) models rigid formats and rhyme with symbols and adapted attention, and ChipSong (Liu et al., 2022) adds word-level lengths, guided attention, and rhyme-oriented training and decoding. DeepRapper (Xue et al., 2021) models rhyme and beats, while Youling (Zhang et al., 2020) supports interactive content and format control. Full-song generation and fine-grained control are also established capabilities: Chae et al. (2025) combine song-form tokens with word-, phrase-, line-, and paragraph-level syllable controls, while leaving genre tags and rhyme schemes to future work in their English setup. Related poetry systems use length and rhyme descriptors learned from non-poetic text with filtering and reranking (Ormazabal et al., 2022), character-level vocabulary restrictions (Yu et al., 2024), constraint-checked search and revision (Zou, 2025), or a rule-based critic of structure, tone, and rhyme for training-data selection (Qu et al., 2026). A complementary line studies melody-conditioned text–music alignment (Watanabe et al., 2018; Sheng et al., 2021), lyric planning with melody-derived constraints at inference (Tian et al., 2023), joint lyric–melody generation (Ding

et al., 2025), and rewriting for a supplied melody (Zhao et al., 2025). LyriCo studies joint Chinese character-group, rhyme, genre, and section targets from a textual brief, using control-token loss weighting and line-local attention bias. Its section annotations provide training supervision rather than a melody input, and text-level adherence does not establish musical alignment. Section 4.4 compares shared constraints with SongNet and ChipSong.

### 2.2 Controllable Text Generation

Controllable text generation spans model adaptation, probabilistic steering, and explicit constrained decoding. Control codes (Keskar et al., 2019), prefix and prompt tuning (Li & Liang, 2021; Lester et al., 2021), and low-rank adaptation (Hu et al., 2022) condition or specialize pretrained models; InstructCTG (Zhou et al., 2023b) learns automatically extracted constraints expressed as natural-language instructions. Generation can also be steered with attribute predictors or contrasting language models (Dathathri et al., 2020; Krause et al., 2021; Yang & Klein, 2021; Liu et al., 2021), energy-based sampling (Qin et al., 2022), or diffusion (Li et al., 2022). Explicit decoding methods instead enforce lexical requirements through grid beam search (Hokamp & Liu, 2017), logical constraints (Lu et al., 2021), and lookahead heuristics (Lu et al., 2022); Ctrl-G (Zhang et al., 2024) combines a distilled hidden Markov model with finite-state constraints. DOMINO (Beurer-Kellner et al., 2024) aligns formal constraints with subword tokenization, and XGrammar (Dong et al., 2025) accelerates grammar checks. Grammar-Aligned Decoding (Park et al., 2024) examines distortion of grammar-conditioned sampling distributions, while sequential Monte Carlo (Loula et al., 2025) uses weighted particles and resampling for syntactic and semantic control. LyriCo uses control tokens and supervised autoregressive learning: its ablations ask whether the two mechanisms improve adherence beyond standard supervised fine-tuning on fixed data and representation.

### 2.3 Constraint Evaluation and Annotation

Constraint evaluation and annotation determine what these control results establish. Numerical planning evaluations (Sun et al., 2023) and COLLIE (Yao et al., 2024) examine constraints across textual scales; FollowBench (Jiang et al., 2024) increases constraint difficulty, and ComplexBench (Wen et al., 2024) evaluates constraint composition. IFEval (Zhou et al., 2023a) distinguishes instruction-level from prompt-level success. These distinctions motivate reporting each metric's evaluation unit: LyriCo's per-metric scores are not all-constraint success rates or directly comparable to these benchmarks. Annotation resources provide a separate source of supervision: Jieba (Sun, 2012) and PaddleSpeech (Zhang et al., 2022) offer lexical and speech-processing infrastructure, while All-In-One (Kim & Nam, 2023) estimates metrical and functional music structure from audio demixed with Hybrid Transformer Demucs (Rouard et al., 2023). Our annotation audits assess labels and boundaries separately from downstream structural adherence (Appendix A.1); annotation quality and generated-text control therefore remain distinct evaluation targets.

## 3 Method

### 3.1 Problem Setting

The input comprises content requirements in a brief $\mathcal{C} = (t, g)$, specifying a theme $t$ and genre $g$, and structural constraints in a brief $\mathcal{S} = (n, \mathbf{p}, \mathbf{l}, \mathbf{f}, \mathbf{r})$. Here, $n$ is the number of lyric lines and $\mathbf{p} = ((s_1, k_1), \ldots, (s_m, k_m))$ specifies section types (e.g., verse and chorus) and their line counts, with $\sum_{j=1}^{m} k_j = n$. For line $i$, $l_i$ is the required Chinese-character count, $\mathbf{f}_i = (f_{i,1}, \ldots, f_{i,u_i})$ is a partition satisfying $\sum_{v=1}^{u_i} f_{i,v} = l_i$ and $f_{i,v} \in \{1, \ldots, 7\}$, and $r_i$ is a requested rhyme class. The groups in $\mathbf{f}_i$ define the *word format*: contiguous spans of the prescribed lengths, separated in the output. These are phrasing units, not necessarily lexical words.

We seek lyrics $\mathbf{y}$ that express $\mathcal{C}$ while satisfying every constraint in $\mathcal{S}$. Chinese-character count is the operational measure of line length; we retain the term *syllable accuracy* when comparing with earlier work. The present setting provides no melody as input and covers Mandarin lyrics with Pop and *gufeng* genre labels.

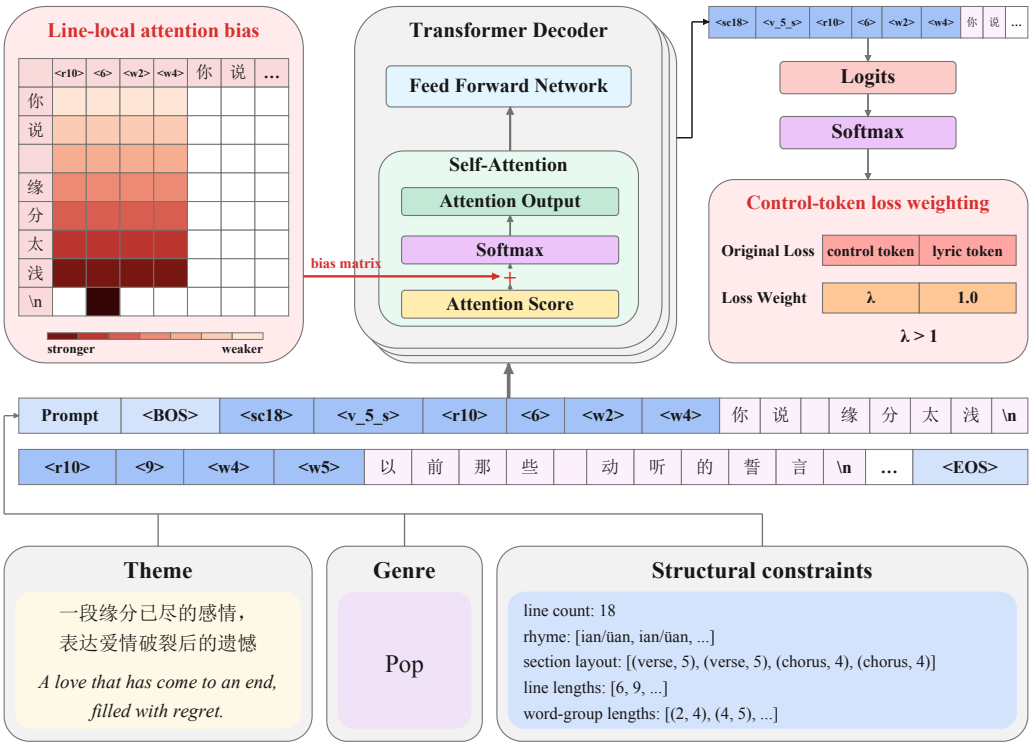


Figure 2: LyriCo architecture. A decoder-only language model generates Chinese lyrics from content requirements and structural constraints encoded as control tokens. Line-local attention bias increases attention to relevant control tokens before softmax, while control-token loss weighting emphasizes their prediction during training. The bottom row illustrates the theme, genre, and structural specifications; English translations are provided for reference.

### 3.2 Structured Representation and Data

We serialize $\mathcal{C}$ and $\mathcal{S}$ into a prompt $\mathbf{x}$ for an autoregressive decoder $p_\theta(\mathbf{y} \mid \mathbf{x})$. The tokenizer is extended with control tokens for sections, line boundaries, line length, rhyme, and word format. Theme and genre remain natural-language fields. Training targets interleave lyric text with control tokens.

**Model computation.** Figure 2 shows the decoder's attention and feed-forward computation alongside the two control mechanisms. Teacher-forced training compares next-token logits with shifted targets; generation appends sampled tokens. Line-local attention bias modifies eligible pre-softmax scores in both stages; loss weighting modifies control-token prediction errors only during training.

**Data construction.** Starting from approximately 250k Mandarin songs paired with audio, the annotation and filtering pipeline retains approximately 190k songs. Classifier-based genre annotation uses 10k manually labeled examples to train a Pop/*gufeng* classifier. A large language model extracts theme labels and checks their consistency with the lyrics. Music structure analysis and lyric alignment derive section annotations by matching predicted boundaries to line-level LRC timestamps. Appendix A.1 describes the annotation steps and structural filters.

For word-format annotation, *TTS-inspired Chinese word segmentation* represents phrase-level units for vocal phrasing and melody planning. Jieba is a lexical word segmenter: 95.0% of its groups contain one or two characters (mean 1.6; Table 1), dividing lyrics into many small lexical units. Lyric lines are typically organized into a few phrases separated by rhythmic pauses or breathing points. Our segmenter derives boundaries from the hierarchy of prosodic words, prosodic phrases, and intonation phrases. Phrase and intonation boundaries mark pauses and potential breathing points.

For example, “张口欲唱 | 声却哑” (“I open my mouth to sing, but my voice fails”) preserves two clauses: the boundary after “唱” provides a natural melodic caesura and breathing point. Jieba instead produces “张口 | 欲 | 唱声 | 却 | 哑”, merging “唱声” across that boundary and making it unavailable for explicit control. TTS-inspired segmentation thus represents the pause and breathing structure needed for vocal phrasing, rather than simply producing longer groups.

An annotation audit obtains 1,592/1,658 (96.02%) exact matches at the lyric-line level, requiring every predicted group boundary to match the audit reference. Additional checks find 100/100 Pop and 98/100 *gufeng* genre labels correct, 100/100 theme labels in an additional random sample judged consistent by GPT, and 398/400 section annotations correct across 100 songs (99/100 whole-song matches). Appendix A.1 gives denominators and criteria.

Table 1: Group-length distributions on the lyrics corpus. TTS-inspired Chinese word segmentation yields coarser groups than Jieba: $87.12\%$ have 3–6 characters. Means are in characters.

| Segmenter | Group length (% of groups) | | | | | | | Mean |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
| jieba | 45.43 | 49.59 | 3.02 | 1.93 | | $<0.03$ | | 1.6 |
| TTS-inspired (ours) | 3.18 | 8.48 | 26.76 | 30.75 | 19.13 | 10.48 | 1.22 | 3.9 |

### 3.3 Control-Token Loss Weighting

Let $\mathcal{T}_{\text{ctrl}}$ contain the control-token positions in a target of valid length $T$. We minimize

$$\mathcal{L}(\theta) = -\frac{1}{T}\sum_{t=1}^{T} w_t \log p_\theta(y_t \mid y_{<t}, \mathbf{x}), \qquad w_t = \begin{cases} \lambda, & t \in \mathcal{T}_{\text{ctrl}}, \\ 1, & \text{otherwise.} \end{cases} \tag{1}$$

The setting $\lambda = 1$ recovers the unweighted objective; $\lambda > 1$ increases the relative contribution of errors on control tokens. Lyric tokens retain unit weight. Control-token loss weighting introduces no trainable parameters.

### 3.4 Line-Local Attention Bias

Line-local attention bias favors relevant control tokens while processing lyric content. We add a bias for eligible query–key pairs within the same lyric line. Let $\mathcal{K}_r$, $\mathcal{K}_\ell$, and $\mathcal{K}_f$ denote positions of rhyme, line-length, and word-format control tokens, respectively. Let $\mathcal{Q}_c$ contain ordinary content-token positions in the target region and $\mathcal{Q}_{\text{nl}}$ contain newline-token positions there. The eligibility indicator is

$$M_{ij} = \mathbf{1}[L(i) = L(j)] \begin{cases} \mathbf{1}[j \in \mathcal{K}_r \cup \mathcal{K}_\ell \cup \mathcal{K}_f], & i \in \mathcal{Q}_c, \\ \mathbf{1}[j \in \mathcal{K}_\ell], & i \in \mathcal{Q}_{\text{nl}}, \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

where $L(i)$ is the line index associated with position $i$. Thus content queries receive a bias toward all three control-token types, while newline queries receive it only toward line-length control tokens. Other control tokens, such as section labels, are excluded. The attention rule is

$$\alpha_{ij} = \text{softmax}_j\left(\frac{\mathbf{q}_i^\top \mathbf{k}_j}{\sqrt{d}} + b_{ij} + a_{ij}\right), \tag{3}$$

$$b_{ij} = \beta M_{ij} \min\left(\gamma_{\max}, \frac{p(i)}{L_{\text{norm}}}\right), \tag{4}$$

where $a_{ij}$ is the usual causal mask, equal to $-\infty$ for future positions. The position $p(i)$ is the one-indexed token position within the line, counting from the first token after the preceding newline. $L_{\text{norm}} > 0$ normalizes position and $\gamma_{\max} > 0$ caps the increase. The design gives later eligible queries a larger preference for visible control tokens, up to this cap.

The same bias rule is used across attention heads during training and generation and adds no trainable parameters. Setting $\beta = 0$ disables it. Section 4.3 varies $\lambda$ and $\beta$; Appendix A.3 lists the training and decoding settings.

Table 2: Functional comparison with prior controllable-lyrics systems. "–" means the control type is not supported; ranges are inclusive. Max length is in lines and word-format range is in characters per group.

| | Genre | Theme | Section | Sentence num range | Word-format | Rhyme | Syllable count |
|---|---|---|---|---|---|---|---|
| SongNet (Li et al., 2020) | – | – | – | $\leq 16$ | – | ✓ | ✓ |
| ChipSong (Liu et al., 2022) | – | only trigger word | – | $\leq 16$ | 1–5 | ✓ | ✓ |
| **Ours** | ✓ | ✓ | ✓ | $\mathbf{\leq 36}$ | **1–7** | ✓ | ✓ |

### 3.5 Inference

At inference time, we serialize $\mathcal{C}$ and $\mathcal{S}$ in the training format and sample autoregressively with top-$p = 0.95$, temperature $0.9$, and the same line-local attention bias. We use neither constrained decoding nor post-hoc rewriting. Constraint adherence is learned and evaluated on generated lyrics, without a hard satisfaction guarantee.

## 4 Experiments

### 4.1 Setup

**Benchmark and models.** The benchmark contains 500 briefs, evenly split between Pop and *gufeng*, varying all five structural constraints. We use DeepSeek-R1-Distill-Qwen backbones at 1.5B and 14B (DeepSeek-AI, 2025). Baselines are standard SFT at 1.5B (L1A0), the unfine-tuned 14B backbone, GPT-5.5 (OpenAI, 2026), and DeepSeek-V4-Pro (DeepSeek-AI, 2026). Further evaluations cover GLM transfer and released ChipSong cases.

**Input and training protocol.** All systems receive the same underlying content and structural requirements. LyriCo uses control tokens; the prompted baselines receive detailed line-by-line instructions with formatting and rhyme examples, but no complete-song exemplar. This comparison evaluates complete adaptation against prompting; the 1.5B ablations fix the backbone, representation, and training recipe while varying the mechanisms. Main models use $\lambda = 3$, $\beta = 3$ (L3A3). Appendix A.3 gives training and decoding settings.

**Metrics.** **SenNumA** is line-count accuracy; **SegA** is section accuracy; **SA** is line-length accuracy, based on Chinese-character counts; **RA** is rhyme accuracy; and **WA** is word-format accuracy, based on requested word-group lengths. **WA-$k$** restricts WA to $k$-character groups. Scores evaluate individual constraints rather than whole-song success (Appendix A.4). Table 3(c) reports Distinct-$n$-based diversity (Li et al., 2016), computed locally and across the corpus, and PPL on 100 held-out lyrics.

Table 2 compares explicit controls and reported ranges in specialized lyric systems. LyriCo combines genre and theme conditioning with section-layout control.

### 4.2 Main Results

Table 3(a) reveals a distinction between global structure and fine-grained control. The API baselines follow section layouts and line counts closely, but are less accurate on line lengths and word-group lengths. LyriCo-14B combines strong global compliance with the highest line-length, rhyme, and word-format accuracy. The main advantage is precise control within individual lyric lines.

The group-length breakdown sharpens this distinction (Table 3(b)). The API baselines handle short word groups well, but their accuracy becomes less consistent for longer groups. LyriCo-14B maintains high accuracy across the full range, with a similar pattern at 1.5B. This stability extends word-format control to longer partitions followed less reliably by the prompted baselines. Figure 3 illustrates these controls in generated lyrics.

The 14B baseline pair shows the effect of complete adaptation: every structural metric improves over the unfine-tuned backbone. At 1.5B, standard SFT already achieves strong line-length and

word-format accuracy (Table 4), showing that learning the structured representation provides a useful foundation. The ablations below examine how emphasizing control tokens improves on this foundation with the backbone and representation held fixed.

Diversity follows a different pattern at local and corpus scales (Table 3(c)). LyriCo-14B leads on corpus-level diversity, while the API baselines lead on local diversity. Strong structural compliance thus coexists with lexical variety across generated lyrics. Lower held-out PPL relative to the same backbone also indicates a closer fit to the lyric distribution.

### 4.3 Ablation of the Two Mechanisms

Across the seven configurations in Table 4, control-token loss weighting provides the clearest gains in line count and section accuracy. This builds on the line-length and word-group structure learned through standard SFT, supporting stronger supervision on lyric-layout specifications.

Moderate line-local attention bias also improves structural accuracy, while a stronger bias offers no consistent additional benefit. The combined configuration achieves the strongest section, line-length, and word-format accuracy and ties for the best rhyme accuracy. Loss weighting alone retains the highest line-count accuracy, so combining the mechanisms balances their contributions across constraints. Improvements over standard SFT extend across all seven word-group lengths. Loss weighting alone gives the highest accuracy for five-character groups, while the combined configuration leads at the other lengths (Table 4).

Table 3: Results on 500 briefs (%), with PPL on 100 held-out lyrics. R1-Qwen = DeepSeek-R1-Distill-Qwen. The 14B and GLM baselines are unfine-tuned. + LyriCo denotes L3A3. Bold marks column maxima, or the minimum PPL; a dash denotes an unreported result.

*(a) Overall structural metrics*

| Model | SenNumA | SegA | SA | RA | WA |
|---|---|---|---|---|---|
| GPT-5.5 | **100.00** | **100.00** | 85.63 | 96.72 | 91.91 |
| DeepSeek-V4-Pro | 99.20 | 99.09 | 73.94 | 95.88 | 82.84 |
| R1-Qwen-1.5B + LyriCo | 96.80 | 98.80 | 98.34 | 95.06 | 98.52 |
| R1-Qwen-14B | 24.80 | 35.61 | 11.17 | 11.39 | 13.35 |
| + LyriCo | **100.00** | 99.75 | **99.07** | **99.00** | **99.30** |
| GLM-9B-0414 | 49.00 | 66.85 | 11.25 | 20.29 | 14.11 |
| + LyriCo | 98.60 | 99.63 | 98.87 | 96.64 | 99.15 |

*(b) Word-format accuracy by group length*

| Model | WA-1 | WA-2 | WA-3 | WA-4 | WA-5 | WA-6 | WA-7 |
|---|---|---|---|---|---|---|---|
| GPT-5.5 | **100.00** | **100.00** | 99.23 | 96.79 | 93.99 | 80.02 | 86.09 |
| DeepSeek-V4-Pro | **100.00** | 97.14 | 93.53 | 89.54 | 83.24 | 65.88 | 75.16 |
| R1-Qwen-1.5B + LyriCo | 96.72 | 98.44 | 98.88 | 99.05 | 98.49 | 98.73 | 98.30 |
| R1-Qwen-14B | 6.14 | 20.09 | 11.04 | 21.09 | 10.06 | 7.20 | 13.29 |
| + LyriCo | 99.51 | 99.51 | **99.39** | **99.73** | **99.46** | **99.57** | **98.83** |
| GLM-9B-0414 | 8.77 | 2.75 | 23.84 | 29.62 | 19.59 | 3.04 | 3.58 |
| + LyriCo | 99.34 | 99.35 | 98.97 | 99.30 | 99.04 | 98.97 | 98.58 |

*(c) Diversity and corpus perplexity*

| Model | MA-D1 | MI-D1 | MA-D2 | MI-D2 | PPL↓ |
|---|---|---|---|---|---|
| GPT-5.5 | 88.54 | 1.84 | **99.42** | 38.02 | – |
| DeepSeek-V4-Pro | **88.90** | 2.29 | 99.13 | 41.23 | – |
| R1-Qwen-1.5B + LyriCo | 81.27 | 2.13 | 95.38 | 34.16 | 9.5942 |
| R1-Qwen-14B | 86.71 | 1.11 | 97.37 | 23.26 | 105.22 |
| + LyriCo | 86.46 | **2.97** | 97.97 | **44.72** | **7.1224** |
| GLM-9B-0414 | 81.79 | 1.93 | 92.34 | 34.29 | 91.85 |
| + LyriCo | 79.44 | 2.43 | 93.27 | 34.64 | 7.18 |

**GUFENG** 《我写春风画秋雨》
Every Stroke is You
对爱人的深情告白
A Deep Confession of Love

| Group | Lyric | Tag |
|---|---|---|
| **VERSE 1** | | rhyme ou / iu |
| 7+3 | 写风话雨道不尽 \| 你温柔 | <10> |
| | Words cannot describe your tenderness | |
| 6+5 | 山河远天长久 \| 思念在心头 | <11> |
| | Mountains and rivers afar, longing stays | |
| **VERSE 2** | | rhyme ao / iao |
| 5+7 | 谁说江湖远 \| 心中事难以预料 | <12> |
| | Who says the world is far and unpredictable | |
| 5+6 | 浮云散去后 \| 方觉世事纷扰 | <11> |
| | Only after clouds fade, we see turmoil | |
| **PRE-CHORUS** | | rhyme i / ü |
| 4+6 | 墨染纸砚 \| 绘出诗情画意 | <10> |
| | Ink paints poetic beauty | |
| 5+7 | 千笔难写下 \| 对你痴情这一句 | <12> |
| | A thousand strokes cannot express love | |
| **CHORUS** | | rhyme i / ü |
| 7+7 | 不恨此生未如愿 \| 只为前世欠佳期 | <14> |
| | No regret for this unfulfilled life | |
| 7+6 | 佛说今生若无缘 \| 来世定会相依 | <13> |
| | If not now, we meet next life | |

**POP** 《进退两难》
Caught in Between
沉溺一段让人窒息的感情
Drowning in a Toxic Love

| Group | Lyric | Tag |
|---|---|---|
| **VERSE 1** | | rhyme i / ü |
| 4+3 | 深夜总是 \| 想起你 | <7> |
| | Late at night, you cross my mind | |
| 4+3 | 却又害怕 \| 失去你 | <7> |
| | Yet so afraid of losing you | |
| **VERSE 2** | | rhyme i / ü |
| 4+3 | 手机屏幕 \| 又亮起 | <7> |
| | The phone lights up again | |
| 4+3 | 困在这场 \| 独角戏 | <7> |
| | Trapped in a solo play | |
| **CHORUS 1** | | rhyme i / ü |
| 4+4+3 | 这种感觉 \| 真的让人 \| 很窒息 | <11> |
| | It truly makes me breathless | |
| 5+5+2 | 不想再维持 \| 这种虚伪的 \| 关系 | <12> |
| | I don't want this fake relationship | |
| **CHORUS 2** | | rhyme i / ü |
| 5+5 | 分不清到底 \| 是爱还是戏 | <10> |
| | I can't tell love from illusion | |
| 4+4 | 我的灵魂 \| 碎了一地 | <8> |
| | My soul shattered | |

Figure 3: Chinese lyric examples in Pop and *gufeng*. Rows show word-group lengths, lyrics, and line-length tags; section headings indicate layout and rhyme. English translations are for reference. Appendix C provides the complete lyrics and control tags.

Table 4: 1.5B ablations on 500 briefs (%). L$x$A$y$ denotes loss weight $\lambda = x$ and attention-bias scale $\beta = y$; L1A0 is standard supervised fine-tuning. Bold marks column maxima, including ties.

| | *Overall structural metrics* | | | | | *Word-format accuracy by group length* | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | SenNumA | SegA | SA | RA | WA | WA-1 | WA-2 | WA-3 | WA-4 | WA-5 | WA-6 | WA-7 |
| L1A0 | 93.40 | 95.19 | 97.58 | 94.26 | 97.96 | 94.72 | 97.54 | 97.88 | 98.13 | 97.65 | 97.77 | 97.62 |
| L2A0 | 95.20 | 97.93 | 97.55 | **95.06** | 97.93 | 95.40 | 98.19 | 98.36 | 98.46 | 97.92 | 98.10 | 97.94 |
| L3A0 | **97.40** | 98.38 | 98.29 | 94.96 | 98.51 | 95.72 | 98.36 | 98.84 | 98.92 | **98.54** | 98.50 | 98.18 |
| L1A1 | 95.80 | 96.80 | 97.70 | 94.77 | 97.97 | 95.37 | 97.37 | 98.07 | 98.29 | 98.07 | 98.13 | 97.71 |
| L1A3 | 95.40 | 96.51 | 98.23 | 94.99 | 98.32 | 96.03 | 98.20 | 98.42 | 98.81 | 98.42 | 98.25 | 97.97 |
| L1A5 | 95.20 | 96.76 | 97.87 | 94.53 | 98.01 | 94.51 | 98.03 | 98.21 | 98.40 | 98.35 | 98.14 | 97.75 |
| L3A3 | 96.80 | **98.80** | **98.34** | **95.06** | **98.52** | **96.72** | **98.44** | **98.88** | **99.05** | 98.49 | **98.73** | **98.30** |

### 4.4 Comparison with Specialized Lyric Systems

We evaluate LyriCo-14B on 506 released ChipSong cases across 20 templates, using the shared line-length, rhyme, and word-format metrics (Table 5). These external templates test control beyond our benchmark. LyriCo maintains high line-length and rhyme accuracy, with strong word-format accuracy across group lengths. The largest difference from the published reference scores is in word-group control, while rhyme accuracy is comparable to ChipSong. SongNet and ChipSong scores are from Liu et al. (2022), Table 1, rather than reruns; Appendix B.4 details the differences in evaluation sets and scoring protocols.

Table 5: LyriCo-14B on 506 released ChipSong cases, alongside published SongNet/ChipSong reference scores from Liu et al. (2022), Table 1. Evaluation protocols differ (Appendix B.4). A dash denotes an unreported result. Scores are percentages; bold marks column maxima.

| Model | SA | RA | WA | WA-1 | WA-2 | WA-3 | WA-4 | WA-5 |
|---|---|---|---|---|---|---|---|---|
| SongNet | 97.80 | 13.82 | 5.89 | 3.67 | 6.82 | 0.75 | 0.59 | – |
| ChipSong | 98.54 | **98.56** | 86.64 | 86.04 | 89.03 | 78.74 | 76.11 | 18.42 |
| LyriCo-14B | **99.83** | 98.42 | **99.93** | **99.92** | **99.96** | **99.70** | **99.51** | **100.00** |

Table 6: GPT-5.6 content-adherence scores (%; 250 outputs/cell). Model notation follows Table 3. Near-ceiling scores limit discrimination among systems.

| | Genre | | Theme | |
|---|---|---|---|---|
| Model | Pop | Gufeng | Pop | Gufeng |
| GPT-5.5 | 100.0 | 100.0 | 100.0 | 100.0 |
| DeepSeek-V4-Pro | 100.0 | 100.0 | 100.0 | 99.6 |
| R1-Qwen-14B | 100.0 | 98.8 | 100.0 | 96.4 |
| + LyriCo | 100.0 | 99.6 | 100.0 | 98.4 |

### 4.5 Transfer and Content Adherence

The complete L3A3 recipe transfers to GLM-9B-0414 on the same 500 briefs and scoring protocol. Every structural metric improves, including word-format accuracy across all seven group lengths (Table 3(a,b)). Consistent control across group lengths suggests that the structured representation remains useful when the underlying language model changes. This experiment tests the portability of the complete framework, complementing the within-backbone ablations that distinguish the two mechanisms. Adaptation also lowers held-out PPL and increases corpus-level diversity, while local unigram diversity decreases (Table 3(c)). As with the 14B model, structural gains accompany broader variation across outputs, with different effects at the local scale.

A GPT-5.6 judge assesses genre and theme adherence for 500 outputs per system (Table 6). All systems score highly on Pop; differences appear in *gufeng*, where LyriCo improves over its backbone on both attributes. These results indicate that structural control remains compatible with high-level content adherence. Since all systems already follow genre and theme closely, the main practical gain is more precise structural control while retaining these content requirements. We assess these attributes separately from artistic quality and singability; the judging protocol is in Appendix B.1.

## 5 Conclusion

We present LyriCo, a structured-text framework for controllable Chinese lyric generation. By encoding structural specifications as control tokens, the framework makes the desired lyric form part of the sequence learned through supervised fine-tuning. Our annotation pipeline supplies the supervision needed to learn this representation. Control-token loss weighting strengthens these signals during training, while a line-local attention bias directs attention toward the controls relevant to the current line. The 1.5B ablations show that emphasizing these signals improves structural compliance beyond standard supervised fine-tuning. At 14B, LyriCo follows fine-grained requirements within lyric lines more reliably than the prompted baselines, while preserving strong control over the overall lyric layout. Applying the same framework to GLM-9B also improves structural adherence, supporting its use across different backbone families. The content-adherence results also show that stronger structural control remains compatible with the requested theme and genre. In lyric writing, a line must fit its prescribed form while contributing to the intended content, so both aspects need to be considered when assessing the generated song. Building on our experiments with Mandarin Pop and *gufeng*, we aim to extend LyriCo to more languages and genres and study how its control mechanisms behave across model scales and training runs. We also plan to incorporate melody as an additional condition for lyric generation and explore how structured lyrics can guide subsequent melody composition.

## AI USE STATEMENT

The data pipeline uses a large language model to extract theme labels from lyrics and to check their consistency with the source text, as described in Section 3.2. GPT checks theme labels in the additional annotation audit (Appendix A.1). GPT-5.6 serves as the content-adherence judge for genre and theme (Section 4.5). During manuscript preparation, an AI assistant (Codex) was used to draft and revise text, search and summarize related literature, audit numerical reporting and evaluation code, suggest evaluation improvements, revise scientific figures, and adapt and check the LaTeX format.

## REPRODUCIBILITY STATEMENT

The structured representation, control-token loss weighting, and line-local attention bias are described in Section 3.2, Section 3.3, and Section 3.4. Data filtering and rhyme groups are documented in Appendix A.1 and Appendix A.2; training and decoding settings appear in Appendix A.3. Evaluation definitions and scope are discussed in Appendix A.4; additional content and transfer results appear in Appendix B.

## REFERENCES

Luca Beurer-Kellner, Marc Fischer, and Martin Vechev. Guiding LLMs the right way: Fast, non-invasive constrained generation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 3658–3673. PMLR, 2024. URL https://proceedings.mlr.press/v235/beurer-kellner24a.html.

Yunkee Chae, Eunsik Shin, Suntae Hwang, Seungryeol Paik, and Kyogu Lee. Song form-aware full-song text-to-lyrics generation with multi-level granularity syllable count control. In *Interspeech 2025*, pp. 1273–1277, 2025. doi: 10.21437/Interspeech.2025-1247. URL https://www.isca-archive.org/interspeech_2025/chae25_interspeech.html.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*, 2020.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

DeepSeek-AI. DeepSeek-V4: Towards highly efficient million-token context intelligence, 2026. URL https://arxiv.org/abs/2606.19348.

Shuangrui Ding, Zihan Liu, Xiaoyi Dong, Pan Zhang, Rui Qian, Junhao Huang, Conghui He, Dahua Lin, and Jiaqi Wang. Songcomposer: A large language model for lyric and melody generation in song composition. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7108–7127, 2025.

Yixin Dong, Charlie F. Ruan, Yaxing Cai, Ruihang Lai, Ziyi Xu, Yilong Zhao, and Tianqi Chen. XGrammar: Flexible and efficient structured generation engine for large language models. In *Proceedings of Machine Learning and Systems*, volume 7, 2025. URL https://proceedings.mlsys.org/paper_files/paper/2025/hash/5c20ca4b0b20b0bd2f1d839dc605e70f-Abstract-Conference.html.

Chris Hokamp and Qun Liu. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1535–1546, 2017.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. FollowBench: A multi-level fine-grained constraints following benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4667–4688, 2024. doi: 10.18653/v1/2024.acl-long.257. URL https://aclanthology.org/2024.acl-long.257/.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*, 2019.

Taejun Kim and Juhan Nam. All-in-one metrical and functional structure analysis with neighborhood attentions on demixed audio. In *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2023.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. Gedi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 4929–4952, 2021.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059, 2021.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 110–119, 2016. doi: 10.18653/v1/N16-1014. URL https://aclanthology.org/N16-1014/.

Piji Li, Haisong Zhang, Xiaojiang Liu, and Shuming Shi. Rigid formats controlled text generation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 742–751, 2020.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597, 2021.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B Hashimoto. Diffusion-LM improves controllable text generation. In *Advances in Neural Information Processing Systems*, 2022.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A Smith, and Yejin Choi. Dexperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6691–6706, 2021.

Nayu Liu, Wenjing Han, Guangcan Liu, Peng Zhou, Ran Zhang, Xiaorui Wang, and Huabin Ruan. Chipsong: A controllable lyric generation system for chinese popular song. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*, pp. 85–95, 2022.

João Loula, Benjamin LeBrun, Li Du, Ben Lipkin, Clemente Pasti, Gabriel Grand, Tianyu Liu, Yahya Emara, Marjorie Freedman, Jason Eisner, Ryan Cotterell, Vikash Mansinghka, Alexander K. Lew, Tim Vieira, and Timothy J. O'Donnell. Syntactic and semantic control of large language models via sequential Monte Carlo. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/a2d537e69a6c6638a3630eef835f07de-Abstract-Conference.html.

Ximing Lu, Peter West, Rowan Zellers, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Neurologic decoding: (un)supervised neural text generation with predicate logic constraints. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4288–4299, 2021.

Ximing Lu, Sean Welleck, Peter West, Liwei Jiang, Jungo Kasai, Daniel Khashabi, Ronan Le Bras, Lianhui Qin, Youngjae Yu, Rowan Zellers, Noah A. Smith, and Yejin Choi. NeuroLogic a*esque decoding: Constrained text generation with lookahead heuristics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 780–799. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.naacl-main.57. URL https://aclanthology.org/2022.naacl-main.57/.

OpenAI. GPT-5.5 System Card, 2026. URL https://openai.com/index/gpt-5-5-system-card/.

Aitor Ormazabal, Mikel Artetxe, Manex Agirrezabal, Aitor Soroa, and Eneko Agirre. PoeLM: A meter- and rhyme-controllable language model for unsupervised poetry generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 3655–3670. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-emnlp.268. URL https://aclanthology.org/2022.findings-emnlp.268/.

Kanghee Park, Jiayu Wang, Taylor Berg-Kirkpatrick, Nadia Polikarpova, and Loris D'Antoni. Grammar-aligned decoding. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2bdc2267c3d7d01523e2e17ac0a754f3-Abstract-Conference.html.

Lianhui Qin, Sean Welleck, Daniel Khashabi, and Yejin Choi. COLD decoding: Energy-based constrained text generation with Langevin dynamics. In *Advances in Neural Information Processing Systems*, 2022.

Zhan Qu, Shuzhou Yuan, and Michael Färber. PoeTone: A framework for constrained generation of structured Chinese Songci with LLMs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(39):32764–32772, 2026. doi: 10.1609/aaai.v40i39.40555. URL https://ojs.aaai.org/index.php/AAAI/article/view/40555.

Simon Rouard, Francisco Massa, and Alexandre Défossez. Hybrid transformers for music source separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023. URL https://arxiv.org/abs/2211.08553.

Zhonghao Sheng, Kaitao Song, Xu Tan, Yi Ren, Wei Ye, Shikun Zhang, and Tao Qin. Songmass: Automatic song writing with pre-training and alignment constraint. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 13798–13805, 2021.

Jiao Sun, Yufei Tian, Wangchunshu Zhou, Nan Xu, Qian Hu, Rahul Gupta, John Wieting, Nanyun Peng, and Xuezhe Ma. Evaluating large language models on controlled generation tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3155–3168, 2023. doi: 10.18653/v1/2023.emnlp-main.190. URL https://aclanthology.org/2023.emnlp-main.190/.

Junyi Sun. Jieba: Chinese text segmentation. https://github.com/fxsjy/jieba, 2012.

Yufei Tian, Anjali Narayan-Chen, Shereen Oraby, Alessandra Cervone, Gunnar Sigurdsson, Chenyang Tao, Wenbo Zhao, Yiwen Chen, Tagyoung Chung, Jing Huang, and Nanyun Peng. Unsupervised melody-to-lyrics generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9235–9254. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.513. URL https://aclanthology.org/2023.acl-long.513/.

Kento Watanabe, Yuichiroh Matsubayashi, Satoru Fukayama, Masataka Goto, Kentaro Inui, and Tomoyasu Nakano. A melody-conditioned lyrics language model. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 163–172, 2018.

Bosi Wen, Pei Ke, Xiaotao Gu, Lindong Wu, Hao Huang, Jinfeng Zhou, Wenchuang Li, Binxin Hu, Wendy Gao, Jiaxin Xu, Yiming Liu, Jie Tang, Hongning Wang, and Minlie Huang. Benchmarking complex instruction-following with multiple constraints composition. In *Advances in Neural Information Processing Systems*, volume 37,

2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f8c24b08b96a08ec7a7a975feea7777e-Abstract-Datasets_and_Benchmarks_Track.html.

Lanqing Xue, Kaitao Song, Duocai Wu, Xu Tan, Nevin L Zhang, Tao Qin, Wei-Qiang Zhang, and Tie-Yan Liu. Deeprapper: Neural rap generation with rhyme and rhythm modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 69–81, 2021.

Kevin Yang and Dan Klein. Fudge: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3511–3535, 2021.

Shunyu Yao, Howard Chen, Austin W. Hanjie, Runzhe Yang, and Karthik Narasimhan. COLLIE: Systematic construction of constrained text generation tasks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=kxgSlyirUZ.

Chengyue Yu, Lei Zang, Jiaotuan Wang, Chenyi Zhuang, and Jinjie Gu. CharPoet: A Chinese classical poetry generation system based on token-free LLM. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 315–325. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-demos.30. URL https://aclanthology.org/2024.acl-demos.30/.

Honghua Zhang, Po-Nien Kung, Masahiro Yoshida, Guy Van den Broeck, and Nanyun Peng. Adaptable logical control for large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/d15c16cf5619a2b1606da5fc88e3f1a9-Abstract-Conference.html.

Hui Zhang, Tian Yuan, Junkun Chen, Xintong Li, Renjie Zheng, Yuxin Huang, Xiaojie Chen, Enlei Gong, Zeyu Chen, Xiaoguang Hu, Dianhai Yu, Yanjun Ma, and Liang Huang. Paddlespeech: An easy-to-use all-in-one speech toolkit. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations*, pp. 114–123, 2022.

Rongsheng Zhang, Xiaoxi Mao, Le Li, Lin Jiang, Lin Chen, Zhiwei Hu, Yadong Xi, Changjie Fan, and Minlie Huang. Youling: An AI-assisted lyrics creation system. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 85–91, 2020.

Songyan Zhao, Bingxuan Li, Yufei Tian, and Nanyun Peng. REFFLY: Melody-constrained lyrics editing model. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 11295–11315, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.564. URL https://aclanthology.org/2025.naacl-long.564/.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023a. URL https://arxiv.org/abs/2311.07911.

Wangchunshu Zhou, Yuchen Eleanor Jiang, Ethan Wilcox, Ryan Cotterell, and Mrinmaya Sachan. Controlled text generation with natural language instructions. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 42602–42613. PMLR, 2023b. URL https://proceedings.mlr.press/v202/zhou23g.html.

Xu Zou. BIPro: Zero-shot Chinese poem generation via block inverse prompting constrained generation framework. In *Proceedings of the 63rd Annual Meeting of the Association for Computational*

*Linguistics (Volume 1: Long Papers)*, pp. 1116–1134. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.56. URL https://aclanthology.org/2025.acl-long.56/.

# A Additional Method Details

## A.1 Data Pipeline

**Corpus and classifier-based genre annotation.** The source pool contains approximately 250k Mandarin songs paired with audio. A genre classifier is trained on a manually labeled, balanced seed set of 10k songs: 5k Pop and 5k *gufeng*. Predictions are filtered at a confidence threshold of 0.8. The retained corpus contains approximately 190k songs; the reported rounded genre counts are 137k Pop and 52k *gufeng*. Pop is a broad corpus category, and the experiments do not separately evaluate finer genre distinctions within it.

**LLM-extracted theme labels.** An LLM first describes the situations, events, or imagery in each lyric and the emotions it expresses. A concise theme label is derived from these descriptions. A second pass checks the label against the complete text. Labels below a confidence threshold are excluded from theme-conditioned training, while their songs can remain available for other control dimensions. An initial manual audit found agreement above 90% on 100 songs. A separate GPT assessment found all 100 theme labels consistent with the lyrics in an additional random sample of songs. The two audits are distinct samples and are not pooled.

**TTS-inspired Chinese word segmentation.** A Chinese predictor estimates prosodic structure using the hierarchy of prosodic words, prosodic phrases, and intonation phrases. The resulting groups define the word-format annotations. Table 1 reports their length distribution alongside Jieba. For the TTS-inspired groups, the three- to six-character bins sum to $26.76 + 30.75 + 19.13 + 10.48 = 87.12\%$; the mean and median lengths are 3.9 and 4 characters. The hierarchy makes phrase boundaries available as explicit word-format targets. Section 3.2 illustrates how these units preserve a clause boundary that lexical segmentation crosses.

**Annotation validation.** Table 7 summarizes the four annotation checks. Genre relabelling is checked against human genre labels on 100 randomly sampled songs per genre; the additional theme audit uses GPT to assess a separate random sample of 100 songs. The section audit evaluates both complete songs and individual sections. For word-group boundaries, an exact-match line is counted only when every predicted boundary agrees with the audit reference. These results describe consistency on the audited samples; they do not by themselves compare TTS-inspired Chinese word segmentation with a length-matched baseline or measure alignment to a particular sung performance.

Table 7: Annotation validation. The theme row reports GPT judgments; the additional sample is separate from the earlier $> 90\%$ check.

| Component | Evaluation unit | Correct / total | Accuracy (%) |
|---|---|---|---|
| Genre | Pop songs | 100/100 | 100.00 |
| | Gufeng songs | 98/100 | 98.00 |
| Theme | Additional sampled songs | 100/100 | 100.00 |
| Sections | Whole songs | 99/100 | 99.00 |
| | Individual sections | 398/400 | 99.50 |
| Word groups | Lines, all boundaries correct | 1,592/1,658 | 96.02 |

**Section annotations and filtering.** Music structure analysis uses an audio-based model to predict section boundaries. Lyric alignment maps these boundaries to LRC line-level timestamps; a line crossing a boundary is assigned to the section covering the majority of its duration. Rule-based post-processing adjusts misalignments. Structural filters retain songs with at most 36 lyric lines, at most 18 Chinese characters per line, fewer than seven lines per section, and word groups of at most seven characters. Line-final text is cleaned to remove garbled characters and special symbols before rhyme annotation.

### A.2 RHYME-CLASS MAPPING

Table 8 defines the 16 task-specific rhyme classes. The evaluator extracts the last Chinese character of a generated line and obtains its tone-free Pinyin final with `pypinyin`. The apical vowel after `zh/ch/sh/r/z/c/s` is assigned a separate identifier, `zhi_i`. These classes operationalize adherence to a requested rhyme; they do not measure rhyme aesthetics or resolve every context-dependent pronunciation.

Table 8: Pinyin-final mapping used for rhyme control. Class indices match the control tags.

| Index | Pinyin finals | Index | Pinyin finals |
|---|---|---|---|
| 0 | e, o, uo, io, er | 8 | ao, iao |
| 1 | ie, ue, üe | 9 | an, uan |
| 2 | a, ia, ua | 10 | ian, üan |
| 3 | i, ü | 11 | ang, iang, uang |
| 4 | ei, uei, ui | 12 | in, ing |
| 5 | ai, uai | 13 | en, un, uen, ün |
| 6 | u | 14 | eng, iong, ong, ueng |
| 7 | ou, iu, iou | 15 | `zhi_i` |

### A.3 TRAINING AND INFERENCE SETTINGS

Table 9 summarizes the training and decoding settings for the Qwen-derived 1.5B and 14B models. Both use a global batch size of 128 and three training epochs. The main models use $\lambda = 3$ and $\beta = 3$; the 1.5B ablation varies these two coefficients.

Table 9: Optimization and decoding settings. GPUs are NVIDIA H20 (96 GB).

| Configuration | 1.5B model | 14B model |
|---|---|---|
| *Training* | | |
| Number of GPUs | 2 | 4 |
| Per-device batch size | 8 | 1 |
| Gradient accumulation steps | 8 | 32 |
| Global batch size | 128 | 128 |
| Learning rate | $2.0 \times 10^{-5}$ | |
| Learning-rate schedule | Cosine | |
| Epochs | 3 | |
| *Inference* | | |
| Temperature | 0.9 | |
| Top-$p$ | 0.95 | |

### A.4 EVALUATION DEFINITIONS AND SCOPE

**Evaluation units.** Structural accuracy is reported separately for section layout, line count, line length, word-group lengths, and rhyme. Under aligned-unit scoring, parse failures are excluded from sample-level accuracies, line comparisons use the overlapping requested/generated range, and designated "center sentences" are exempt from some length, rhyme, and word-group checks. Missing groups within an aligned line remain in the word-group denominator. Thus line- and group-level scores condition on the evaluated units; they are not strict success rates over every requested constraint and all 500 briefs. A correct rhyme or word partition alone does not imply the requested line count or section sequence is correct.

**Text and control tokens.** Length and word-format checks concern actual Chinese characters and their group boundaries. Emitting a line-length control token is distinct from producing text of that length. Figures 1–3 use simplified display notation, explained in Appendix C.

**Rhyme.** Rhyme accuracy compares the final Chinese character's tone-free Pinyin final with the requested rhyme class using the mapping in Appendix A.2. Trailing punctuation is ignored. The apical vowel following `zh/ch/sh/r/z/c/s` has a separate class. This tests membership in the specified rhyme group; it does not score rhyme aesthetics or melody-dependent pronunciation.

**Diversity and interpretation.** MA-D1/MA-D2 summarize local lexical diversity, while MI-D1/MI-D2 summarize diversity across the generated collection. Local repetition and collection-wide vocabulary coverage describe different properties, so the two levels are reported separately. PPL characterizes model likelihood on held-out lyrics. These automatic statistics do not measure human preference, thematic originality, or singability, and are interpreted alongside structural adherence rather than as a combined measure of lyric quality.

# B Additional Evaluation Results

## B.1 Content Adherence

The GPT-5.6 content-adherence judge evaluates whether each generated lyric is consistent with its requested genre and theme. Each of the four systems contributes 500 outputs, including 250 Pop and 250 *gufeng* outputs, for 2,000 outputs in total. Table 6 reports the percentage judged consistent for each attribute and genre. These are two separate judgments about high-level content, not assessments of melodic compatibility, originality, or artistic preference.

The fine-tuned model's *gufeng* scores increase from 98.8% to 99.6% for genre and from 96.4% to 98.4% for theme relative to its backbone; both retain 100% on the Pop subset. GPT-5.5 receives 100% in all four cells. Thus these measurements provide complementary evidence that structural gains do not coincide with an obvious loss of high-level content adherence under this judge. A single judge with a strong ceiling effect does not establish equivalent creative quality or rule out subtler content differences.

## B.2 Main-Benchmark Diversity and Perplexity

Table 3(c) complements its structural panels. Local and corpus-level diversity need not rank models identically: LyriCo-14B has the highest MI-D1 and MI-D2, while API models have the highest MA-D1 and MA-D2. PPL is evaluated on 100 held-out lyrics; it describes model likelihood rather than human preference. The GLM rows use the same 500-brief benchmark: MA-D1 falls from 81.79% to 79.44%, while MI-D1, MA-D2, and MI-D2 rise, and PPL decreases from 91.85 to 7.18. All GLM results are included in Table 3; this evaluates transfer of the complete L3A3 recipe.

## B.3 Word-Format Accuracy by Group Length

Table 4 reports word-format accuracy by required group length alongside the overall structural metrics. L3A3 improves WA-1 from 94.72% to 96.72% and WA-7 from 97.62% to 98.30% relative to L1A0. L3A0 has the highest WA-5, so the combined configuration is not best for every group length. These comparisons hold the backbone and training recipe fixed; the larger cross-system gaps in Table 3 reflect the complete adapted systems.

## B.4 Released ChipSong Cases and Published References

LyriCo-14B is evaluated on 506 cases across 20 templates from ChipSong's released evaluation resources. The comparison covers SA, RA, WA, and WA-1–WA-5; theme, genre, and section controls are outside its scope. The SongNet and ChipSong rows reproduce Table 1 of Liu et al. (2022), whose Section 4.2 describes 500 templates and 20 generated outputs per template (10,000 outputs). These rows therefore serve as published reference results, not reruns on the 506 LyriCo cases.

ChipSong defines WA as sentence-level exact word-layout accuracy, and WA-$k$ checks the position and length of individual $k$-character words. Its published rhyme evaluator uses a 14-group scheme. These definitions are distinct from the individual-group WA aggregation and 16 rhyme classes of our main benchmark. Table 5 retains the reported metric labels and scores for the 506-case evaluation and the published reference rows; their effective denominators are not standardized by this comparison.

They should be interpreted as reference results rather than a sample-matched ranking. The unreported SongNet WA-5 result is marked with a dash. The SongNet row is the version evaluated by ChipSong's authors on their corpus, and its WA-4 is 0.59% in the published Table 1.

# C Full Lyric Examples and Display Notation

The complete Chinese lyric examples from Figures 1 and 3 appear below with control tags. English translations are provided for reference. Structural targets apply to Chinese. Section headings identify verses and choruses; genre headings repeat when examples continue across pages.

The drawings use display abbreviations: `<rhyme_k>` denotes rhyme class $k$, `<word_k>` a $k$-character group, and a numeric tag such as `<10>` the displayed line-length label. These conventions are not a specification of the tokenizer or the evaluator's exact input syntax. The Pinyin annotations follow the rhyme-class notation in Table 8.

## C.1 Complete Shared-Constraint Examples

Both ten-line examples from Figure 1 are reproduced in parallel with their control tags. Bars mark word groups; colored final characters mark rhyme. The verse has four lines and the chorus has six, all in rhyme class 3.

在爱情中患得患失、纠结徘徊的心境
The anxious, wavering state of mind in love
– caught between hope and fear, unable to decide.

| **POP** 《该不该》<br>Should I? | | **GUFENG** 《难生恨》<br>Hard to Hate | |
|---|---|---|---|
| **VERSE (4 lines)** | | **VERSE (4 lines)** | |
| <rhyme_3><5><word_2><word_3><br>窗外\|雨在滴<br>Outside, the rain keeps dripping | (i) | <rhyme_3><5><word_2><word_3><br>那日\|听莺啼<br>That day, I heard the orioles' song | (i) |
| <rhyme_3><5><word_2><word_3><br>思念\|没道理<br>This longing defies all reason | (i) | <rhyme_3><5><word_2><word_3><br>添香\|又并蒂<br>Burning incense, twin blooms as one | (i) |
| <rhyme_3><5><word_2><word_3><br>明明\|想见你<br>Clearly, I want to see you | (i) | <rhyme_3><5><word_2><word_3><br>今朝\|闻飞絮<br>This morning, catkins fill the air | (ü) |
| <rhyme_3><5><word_2><word_3><br>害怕\|会失去<br>Yet I'm afraid I might lose you | (ü) | <rhyme_3><5><word_2><word_3><br>零落\|随风起<br>Scattered petals rise with the wind | (i) |
| **CHORUS (6 lines)** | | **CHORUS (6 lines)** | |
| <rhyme_3><10><word_5><word_5><br>到底该不该\|真心掏给你<br>Should I hand you my bare heart? | (i) | <rhyme_3><10><word_5><word_5><br>我无悔相遇\|相知到相惜<br>No regret meeting, knowing, treasuring you | (i) |
| <rhyme_3><10><word_5><word_5><br>还是该保留\|防备当武器<br>Or hold back, guard as a weapon? | (i) | <rhyme_3><10><word_5><word_5><br>只恨道不尽\|忘却又铭记<br>Hate that words fail – forgotten, yet engraved | (i) |
| <rhyme_3><12><word_4><word_4><word_4><br>互相拉扯\|实在压抑\|让人窒息<br>Push-pull, so stifling, it suffocates | (i) | <rhyme_3><12><word_4><word_4><word_4><br>纵使千般\|思绪万缕\|痛彻心底<br>A thousand thoughts, piercing deep | (i) |

在爱情中患得患失、纠结徘徊的心境
The anxious, wavering state of mind in love
– caught between hope and fear, unable to decide.

**POP** 《该不该》
Should I?

<rhyme_3><12><word_4><word_4><word_4>
不断猜疑|无法逃避|失去力气 (i)
Endless doubt, no escape, losing strength

<rhyme_3><10><word_5><word_5>
想要去放弃|却怕错失你 (i)
Want to give up, yet dread losing you

<rhyme_3><8><word_4><word_4>
最后只能|独自哭泣 (i)
In the end, I can only weep alone

**GUFENG** 《难生恨》
Hard to Hate

<rhyme_3><12><word_4><word_4><word_4>
回眸一眼|见君转身|泪眼迷离 (i)
One glance back – you turn – eyes blur

<rhyme_3><10><word_5><word_5>
若爱是场戏|我入得痴迷 (i)
If love's a play, I'm utterly entranced

<rhyme_3><8><word_4><word_4>
尽管早已|注定分离 (i)
Though from the start, parting was foretold

## C.2 Complete Tagged Lyric Examples

The complete *gufeng* and Pop lyrics corresponding to Figure 3 are shown in parallel. Each entry contains its control tags, Chinese lyric, Pinyin final, and reference English translation. Section headings give the line counts.

**GUFENG** 《我写春风画秋雨》
Every Stroke is You
对爱人的深情告白
A Deep Confession of Love

**VERSE 1 (4 lines)**

<rhyme_7><10><word_7><word_3>
写风话雨道不尽|你温柔 (ou)
Words cannot describe your tenderness

<rhyme_7><10><word_7><word_3>
花开花落不回头|我等候 (ou)
Flowers bloom and fade, I still wait

<rhyme_7><11><word_6><word_5>
山河远天长久|思念在心头 (ou)
Mountains and rivers afar, longing stays

<rhyme_7><10><word_7><word_3>
酒醉梦醒终别离|几时休 (iu)
Drunk then awake, when will it end

**VERSE 2 (4 lines)**

<rhyme_8><12><word_5><word_7>
谁说江湖远|心中事难以预料 (iao)
Who says the world is far and unpredictable

<rhyme_8><12><word_5><word_7>
谁说红尘乱|不解人间的喧闹 (ao)
Who says the mortal world is chaotic

**POP** 《进退两难》
Caught in Between
沉溺一段让人窒息的感情
Drowning in a Toxic Love

**VERSE 1 (4 lines)**

<rhyme_3><7><word_4><word_3>
深夜总是|想起你 (i)
Late at night, you cross my mind

<rhyme_3><7><word_4><word_3>
刺痛翻涌|在心底 (i)
Sharp pain surges in my heart

<rhyme_3><7><word_4><word_3>
拼命想要|去抽离 (i)
Trying so hard to pull away

<rhyme_3><7><word_4><word_3>
却又害怕|失去你 (i)
Yet so afraid of losing you

**VERSE 2 (4 lines)**

<rhyme_3><7><word_4><word_3>
手机屏幕|又亮起 (i)
The phone lights up again

<rhyme_3><7><word_4><word_3>
以为是你|发消息 (i)
Thinking it is your message

**GUFENG** 《我写春风画秋雨》
Every Stroke is You
对爱人的深情告白
A Deep Confession of Love

<rhyme_8><11><word_5><word_6>
浮云散去后|方觉世事纷扰 (ao)
Only after clouds fade, we see turmoil

<rhyme_8><12><word_5><word_7>
我愿将心事|随风托与白云飘 (iao)
I entrust thoughts to drifting clouds

**PRE-CHORUS (4 lines)**

<rhyme_3><10><word_4><word_6>
墨染纸砚|绘出诗情画意 (i)
Ink paints poetic beauty

<rhyme_3><10><word_3><word_7>
只为你|提笔万般皆是你 (i)
All I write is for you

<rhyme_3><7><word_7>
不求世人知心意 (i)
I seek no recognition

<rhyme_3><12><word_5><word_7>
千笔难写下|对你痴情这一句 (ü)
A thousand strokes cannot express love

**CHORUS (4 lines)**

<rhyme_3><14><word_7><word_7>
不恨此生未如愿|只为前世欠佳期 (i)
No regret for this unfulfilled life

<rhyme_3><14><word_7><word_7>
来生若遇再续缘|失魂梦断也愿意 (i)
If reborn, I will love you again

<rhyme_3><13><word_7><word_6>
我在佛前叩问着|宿命有何玄机 (i)
I ask fate before Buddha

<rhyme_3><13><word_7><word_6>
佛说今生若无缘|来世定会相依 (i)
If not now, we meet next life

**POP** 《进退两难》
Caught in Between
沉溺一段让人窒息的感情
Drowning in a Toxic Love

<rhyme_3><7><word_4><word_3>
期待瞬间|落谷底 (i)
Hopes instantly crash down

<rhyme_3><7><word_4><word_3>
困在这场|独角戏 (i)
Trapped in a solo play

**CHORUS 1 (5 lines)**

<rhyme_3><10><word_5><word_5>
进退两难的|纠缠太压抑 (i)
This dilemma suffocates me

<rhyme_3><11><word_4><word_4><word_3>
这种感觉|真的让人|很窒息 (i)
It truly makes me breathless

<rhyme_3><11><word_4><word_4><word_3>
像是被谁|牵着鼻子|在演戏 (i)
Like being controlled in a play

<rhyme_3><12><word_5><word_5><word_2>
不想再维持|这种虚伪的|关系 (i)
I don't want this fake relationship

<rhyme_3><10><word_5><word_5>
却又沉溺于|谎言的甜蜜 (i)
Yet I indulge in sweet lie

**CHORUS 2 (5 lines)**

<rhyme_3><10><word_5><word_5>
分不清到底|是爱还是戏 (i)
I can't tell love from illusion

<rhyme_3><10><word_5><word_5>
这种感觉像|吞进了玻璃 (i)
Feels like swallowing glass

<rhyme_3><10><word_4><word_6>
每走一步|都要小心翼翼 (i)
Every step is cautious

<rhyme_3><8><word_4><word_4>
耗尽力气|困在原地 (i)
Drained of strength, stuck in place

<rhyme_3><8><word_4><word_4>
我的灵魂|碎了一地 (i)
My soul shattered