CordelSextilha.BR: A Benchmark for Poetic Form in Brazilian Cordel Verse Generation

Bryan K. S. Barbosa[1,2],
Marcela Y. A. Barbosa[3]

[1] Núcleo Interinstitucional de Linguística Computacional (NILC)

[2]Programa de Pós-Grad. em Linguística, Universidade Federal de São Carlos (UFSCar)
São Carlos/SP – Brazil

[3]Unidade Acadêmica de Letras, Universidade Federal de Campina Grande (UFCG)
Campina Grande/PB – Brazil

bryankhelven@ieee.org, marcela.yara@estudante.ufcg.edu.br

Abstract. *We introduce CordelSextilha.BR, the first benchmark for automatic generation of Brazilian cordel sextilhas. We compile 1,519 public-domain stanzas, remove the final line, and gather expert ratings for rhyme, meter, coherence, and "cordelisticity". Rule-based metrics (RhythmAcc, RhymeAcc) align with human judgements; GPT-4o and LLaMA-3.2 (8B) reach about 80% rhyme and about 75% rhythm accuracy. The corpus, metrics and baselines enable RLHF research in low-resource poetry.*

1. Introduction

Brazil's *literatura de cordel* – rhymed pamphlets traditionally sung or declaimed in the North-East – was inscribed on the National Register of Intangible Cultural Heritage by IPHAN in 2018 [Instituto do Patrimônio Histórico e Artístico Nacional (IPHAN) 2018]. Yet, despite this official recognition and more than two decades of federal safeguarding policy, the digital footprint of *cordel* remains fragmentary: most online collections are unannotated scans or missing links, and there are no publicly available corpora that capture its strict metrical template (seven-syllable *redondilha maior*) or its characteristic ABCB rhyme scheme. This digital absence hampers preservation efforts highlighted by heritage surveys [Martins et al. 2023] and blocks advances in Natural Language Processing (NLP), where low-resource poetic traditions are systematically overlooked.

At the same time, large language models (LLMs) have begun to master poetic form in high-resource languages, assisted by transparent, rule-based metrics and, more recently, Reinforcement Learning from Human Feedback (RLHF). Benchmarks for English sonnets [Walsh et al. 2024], Chinese quatrains [Wang et al. 2016] and Russian stress-verse [Koziev 2025] demonstrate that hard meter-and-rhyme checks correlate strongly with human judgements and can be used as rewards to fine-tune LLMs. No analogous benchmark exists for Portuguese – let alone for the *sextilha*, the six-line stanza that anchors cordel poetics.

This paper fills that gap. We introduce **CordelSextilha.BR**[1], the first open benchmark for evaluating poetic form in Brazilian cordel generation. Our contributions are fourfold:

[1]https://github.com/bryankhelven/CordelSextilhaBR

1. **Corpus**. A collection of 1,519 sextilhas extracted from 18 public-domain *folhetos de cordel*, with the final (sixth) line removed for conditional generation tasks.
2. **Human Annotations**. Two linguists scored each machine-generated line for rhythm, rhyme, coherence and stylistic authenticity, achieving substantial inter-annotator agreement (K = 0.91).
3. **Adapted Metrics**. We release *RhythmAcc* and *RhymeAcc*, rule scripts that extend Portuguese scansion tools [Mittmann et al. 2018, Gonçalo Oliveira et al. 2007] with North-Eastern phonology and ABCB rhyme detection, raising rhythm accuracy from 18% to 75% after linguistic refinements.
4. **Baselines & RLHF Pilot**. We report strong baselines with GPT-4o and LLaMA 3.2 (8B) and show that a small RLHF loop, rewarded by our metrics, further boosts form compliance.

Beyond advancing computational creativity in Portuguese, CordelSextilha.BR illustrates how heritage preservation and NLP can be mutually reinforcing: cultural constraints supply objective evaluation hooks, while model-generated variants paves novel avenues for community engagement and teaching.

The remainder of this paper is structured as follows: Section 2 provides the theoretical background, positioning *cordel* within Brazil's heritage policy and outlining the formal-poetics lens adopted here. Section 3 surveys related NLP work on verse generation and evaluation. Section 4 details our methodology – corpus compilation, annotation protocol, and metric engineering. Section 5 reports automatic and human evaluation results. Section 6 concludes with limitations and future work, including larger-scale RLHF experiments and multimodal archiving plans.

2. Theoretical Background

The present work occupies the confluence of cultural heritage preservation and computational creativity, seeking to bridge the gap between Brazil's rich oral-print tradition and the automated generation and evaluation of poetic forms. To this end, we advance through five interlocking strands of prior research: (1) the patrimonial status of *cordel* literature within Brazilian intangible heritage, (2) its formal poetics and metrical constraints, (3) past endeavors in automatic poetry generation and assessment, (4) the emergence of interpretable evaluation metrics (including reinforcement-learning approaches), and (5) the persistent scarcity of resources tailored to Brazilian Portuguese. By moving systematically from the cultural stakes to the technical lacuna, we establish both the humanistic imperative and the computational opportunity that motivate our study.

Importantly, each strand provides context and also highlights the evolving demands placed on digital humanities and Natural Language Processing (NLP) practitioners. In particular, while substantial work has been devoted to poetry generation in globally dominant languages, the specific metrical forms and community-sensitive evaluation frameworks of *cordel* remain under-explored. The next subsection anchors this argument by examining how *cordel* has been valorized as intangible heritage and how its formal constraints supply a concrete test bed for computational models.

2.1. Cordel as Brazilian Intangible Heritage

Literatura de Cordel is a popular verse genre traditionally distributed in low-cost pamphlets and performed orally across Brazil's North-East region. In

2018, the Brazilian National Institute of Historic and Artistic Heritage (IPHAN) inscribed *cordel* on the national list of Intangible Cultural Heritage, emphasizing its role in safeguarding collective memory and regional identity [Instituto do Patrimônio Histórico e Artístico Nacional (IPHAN) 2018]. Subsequent policy analyses argue that such patrimonialization demands preservation of traditional practices and also encourages their adaptive circulation in contemporary and digital media [Nogueira 2018]. Despite this, digital humanities initiatives remain sparse: extant collections often consist merely of scanned *folhetos* without searchability or linguistic annotation. This lacuna underscores the dual necessity of our project – to serve both humanistic scholarship and technical innovation – and situates our corpus-building endeavor within a broader agenda of cultural preservation and public engagement.

Building on this cultural imperative, we turn to the formal poetics that define *cordel* verse. Canonical *sextilhas* employ *redondilha maior* – lines of exactly seven poetic syllables – and an *ABCB* rhyme scheme, where lines 2 and 4 rhyme and that rhyme is echoed in line 6 [Rigonatto 2025]. Didactic materials used in community workshops further reveal that cordelistas perceive the sixth line as the stanza's syntactic and rhythmic closure, resolving both narrative tension and metrical expectations [Pinheiro 2023]. From a computational perspective, these well-defined constraints offer an objective evaluation framework: any system tasked with generating or completing the sixth line must simultaneously satisfy syllabic count, rhyme agreement, and semantic coherence with the preceding five lines. By foregrounding *redondilha* and *ABCB* rhyme, our evaluation metrics are thus grounded in features that reflect practitioners' aesthetic judgwments, ensuring that machine assessments align closely with cultural norms and community values.

2.2. Conceptual Overview of RLHF for Stylistic Alignment

Reinforcement Learning from Human Feedback (RLHF) couples a *policy* language model with a *reward* model trained on expert preferences. Instead of maximising likelihood of next-token distributions, the policy maximises expected reward $R(y, x)$ given an input x and its generated continuation y. In practice, R is a scalar distilled from pairwise comparisons: annotators rank candidate outputs, and the reward model regresses those rankings. Gradients are then estimated by Proximal Policy optimization (PPO) or a KL-regularised variant that keeps the policy close to its pre-trained prior.

Because R can encode *any* measurable signal, RLHF lends itself to stylistic control. For poetry, a reward may combine RHYTHMACC, RHYMEACC, and semantic coherence, thereby steering generations toward prosodic fidelity while preserving narrative sense. Unlike hard post-filters, RLHF integrates constraints into the decoding distribution itself, reducing degeneration and mode-collapse. Recent LLM studies show that even a small set ($\sim$500 examples) of high-quality comparisons can significantly improve form compliance on out-of-domain prompts [Chen et al. 2024].

2.3. Research Question and Motivation Bridge

The cultural stakes of safeguarding *cordel* and the strict formalism of the *sextilha* jointly raise the following research question:

> *Can a compact Portuguese LLM generate the **sixth line** of a sextilha while simultaneously (i) maintaining the seven-syllable redondilha maior, (ii)*

matching the ABCB rhyme nucleus, and (iii) producing a semantically coherent closure?

Answering this question requires three assets that do not yet exist for Brazilian Portuguese: (1) a sizeable, annotated *sextilha* corpus; (2) transparent, rule-based evaluation metrics aligned with cordelistas' own criteria; and (3) an optimization scheme – here, RLHF – that can learn from a handful of expert comparisons. The remainder of the paper details how we build these assets and empirically test their interplay.

3. Related Work

The literature that underpins this study bridges cultural heritage, formal poetics, and the rapidly growing field of large-scale language modeling. We review six strands of work in a cumulative fashion: each subsection motivates the next, moving from the cultural legitimacy of *cordel* to the most recent reinforcement-learning approaches to poetic form.

Globally, several digital heritage platforms and standards have been developed to support documentation of cultural assets. For example, the Arches Project (2010–2011) is an open-source, GIS-based data management system jointly developed by the Getty Conservation Institute and the World Monuments Fund. Arches provides templates and workflows for inventorying both tangible and intangible heritage, and relies on semantic standards (such as CIDOC-CRM) to structure the information [Araujo et al. 2019]. Similarly, the MIDAS-Heritage guidelines (Forum on Information Standards in Heritage, UK) define best practices and data fields for recording historical sites and cultural objects [Araujo et al. 2019]. In general, these systems demonstrate how geospatial technologies and metadata standards can be combined to create integrated heritage archives.

In the Brazilian context, a few notable projects have sought to apply such frameworks. For instance, IPHAN's own Sistema Integrado de Conhecimento e Gestão (SICG, launched in 2012) was an early initiative to geolocate and manage heritage inventories nationwide [Araujo et al. 2019]. Other research has applied photogrammetry and remote sensing to record architectural heritage [Campos et al. 2015], suggesting analogous approaches could capture intangible heritage sites or artifacts. Nonetheless, recent surveys of digital cultural heritage in Brazil reveal significant challenges: many collections are fragmented across institutions, with inconsistent metadata and limited online access [Martins et al. 2023]. Martins' work report that Brazilian cultural datasets often lack systematized organization and open licensing, which hampers data reuse and synthesis [Martins et al. 2023]. This situation indicates a need for improved data infrastructure and interoperability in Brazil's heritage sector.

Academic work on digital heritage in Brazil has thus far focused more on tangible culture and conservation workflows. Araujo *et al.* (2019) systematized Brazilian research on digital tools for architectural documentation, finding growing interest in 3D modeling, BIM, GIS and data standards [Araujo et al. 2019]. In contrast, studies specifically targeting the digitization of living traditions or intangible practices are scarce. There is emerging interest in novel technologies – for example, the use of AR/VR and online archives to share cultural performances – but practical implementations are few. Future work may adapt these international models (Arches, CIDOC-CRM, etc.) and national policies to create digital platforms tailored for Brazil's intangible heritage communities.

In summary, there exists a foundation of international standards and tools for digital heritage documentation, along with Brazilian initiatives in inventorying and technology application. However, the particular requirements of intangible heritage (community involvement, multimedia content, dynamic practices) demand specialized attention. The institutional framework (IPHAN, laws, inventory programs) provides justification and guidance for such efforts. Building on this context and prior work, subsequent sections will outline our methodological approach to developing a digital heritage system aligned with Brazilian cultural policy and technological practices.

3.1. Form-Compliance Evaluation in Large Language Models

Early work on quatrains and pentameter suggested that hard-coded rule checks correlate with perceived quality (as in subsection 3.2). This hypothesis is now confirmed at scale: [Walsh et al. 2024] introduce *Sonnet or Not, Bot?*, a binary form-accuracy task for English sonnets that yields a Pearson correlation of 0.72 with crowd judgwments on fluency and aesthetics. In a different linguistic setting, [Koziev 2025] release RIFMA, a stress-marked corpus for Russian verse, and show that rhyme+meter accuracy explains over 60% of the variance in expert scores. Together, these studies indicate that simple, language-specific rule metrics can serve as transparent proxies for human evaluation – an insight we operationalise through *RhythmAcc* and *RhymeAcc* for Portuguese *sextilhas*.

3.2. Automatic Poetry Generation and Last-Line Tasks

Fixed-form constraints have catalysed benchmark design in other poetic traditions. Wang *et al.* pioneered a planning-based RNN that generates the final line of classical Chinese quatrains, demonstrating that treating the closing line as a conditional generation problem yields coherent poems that satisfy tonal rules [Wang et al. 2016]. Greene *et al.* introduced automatic scansion for English iambic pentameter, linking rhythmic correctness to perceived fluency [Greene et al. 2010]. These studies collectively validate the "last-line" paradigm: the model is judged on its ability to complete a partially given stanza while obeying hard formal constraints. By analogy, generating the sixth line of a *sextilha* requires a model to infer the rhyming nucleus from lines 2 and 4 and to maintain seven-syllable rhythm – an arguably stricter test because rhyme and meter must co-occur.

3.3. Interpretable Metrics and Explainable AI

Explainable AI (XAI) advocates strongly emphasise the need for feedback mechanisms that domain experts can directly inspect and understand. This approach seeks to enhance the interpretability of computational outputs while building trust among users who rely on these technologies. For example, [Yakovenko 2020] from Facebook AI Research (FAIR) conceptualised rhythmic verse generation as a constraint-satisfaction problem, explicitly revealing phoneme-level patterns to the user.

Extending this principle beyond English poetry, recent research by [Al-Rashid and Ali 2025] utilised Bi-directional Long Short-Term Memory (Bi-LSTM) neural networks to classify the sixteen quantitative meters of classical Arabic poetry. Their model achieved an impressive accuracy rate of 97%, coupled with visual highlighting of metrical violations. This visual feedback made errors easily interpretable, facilitating both understanding and correction by poets. Similarly, Facebook AI's rhythmic evaluation toolkit for rap lyrics provided intelligible, diagnostic feedback on generated verses [Yakovenko 2020].

3.4. Portuguese Scansion & Rhyme Tools

Aoidos. *Aoidos* is a rule-based system that automatically performs syllabic scansion and rhyme classification for Portuguese and Spanish verse, achieving per-syllable accuracy above 90% on Camões-era texts [Mittmann et al. 2018]. Although Aoidos can count syllables, its heuristic for synalepha treats the hiatus differently from popular *cordel* pronunciation, yielding systematic off-by-one errors in *redondilha maior*. Moreover, its rhyme detector assumes final-word stress, whereas ABCB rhyme in *cordel* often employs proparoxytone endings.

SilabasPT3 & ancillary APIs. SilabasPT3, released as part of the *Tra-la-Lyrics* project, exposes an API for syllable segmentation and stress position identification, with downstream use in rhythm-aware lyric generation [Gonçalo Oliveira et al. 2007]. While useful for generic verse, SilabasPT3 returns token-level counts only, lacking stanza-level rhyme schema detection and therefore cannot evaluate ABCB compliance.

PoeTree treebanks. The POETREE corpus aggregates $\sim$330k poems across ten languages, including 28k Portuguese texts enriched with Universal Dependencies and basic stanza segmentation [Plecháč et al. 2024]. However, its Portuguese subset is dominated by 19th–20th century lyric poetry; meter/rhyme annotations are absent, and no poem follows the *sextilha* template. Consequently, PoeTree is ill-suited as a benchmark for cordel-specific form metrics.

Gap. No existing tool explicitly targets (i) seven-syllable scansion under Brazilian North-Eastern phonology or (ii) automatic detection of the "2–4–6" ABCB rhyme pivot. Our work therefore introduces bespoke rule scripts that extend Aoidos' algorithm with cordel-specific diacritic normalization and diphthong handling, filling this methodological gap.

3.5. Multilingual Benchmarks for Poetic Structure

Form-centric evaluation is not confined to Portuguese. In Chinese, [Wang et al. 2016] propose a keyword-planned encoder–decoder that generates the final line of regulated quatrains; their dataset of 2,760 quatrains constitutes the first open benchmark with tonal annotations. For English, the *Sonnet or Not, Bot?* corpus provides 7,200 human-rated sonnets with binary rhyme & iambic pentameter labels [Walsh et al. 2024]. Russian is covered by the RIFMA treebank, pairing 4,500 stanzas with stress patterns and phonetic rhyme strings [Koziev 2025]. Recently, a Frontiers study released an Arabic dataset spanning the sixteen classical meters and validated a Bi-LSTM meter classifier at 97% accuracy [Al-Rashid and Ali 2025].

These resources confirm that simple rule checks scale across languages, yet none addresses Brazilian Portuguese, nor the ABCB *sextilha* pattern, underscoring the need for the benchmark we introduce here.

3.6. Reinforcement Learning for Poetic Form

While static metrics can diagnose errors, they do not prescribe how to correct them. Reinforcement Learning from Human Feedback (RLHF) has emerged as a pragmatic

avenue for aligning LLM outputs with non-trivial stylistic constraints. Recent arXiv studies fine-tune small models using composite rewards that combine meter, rhyme and semantic coherence, obtaining double-digit accuracy gains over supervised baselines [Chen et al. 2024]. Integrating RLHF into the evaluation loop thus offers a path from assessment to improvement, while keeping the optimization objective interpretable. We adopt this strategy in a pilot study – *RLHF-Rhythm* – to show that modest amounts of annotated *sextilhas* suffice to steer a compact model toward higher form compliance.

3.7. Summary and Remaining Gap

Taken together, prior research establishes that: (i) form-aware poetry generation tasks are legitimate NLP benchmarks; (ii) rule-based metrics correlate well with human assessment; and (iii) RLHF can enhance form obedience when guided by transparent rewards. However, all existing resources focus on Chinese, English, Russian or Arabic traditions, leaving Brazilian Portuguese – and *cordel* in particular – without corpora, metrics or baselines. The present work fills this void by releasing a sizeable *sextilha* corpus, proposing redondilha- and rhyme-centric metrics, validating them against expert annotation, and demonstrating RLHF gains. In doing so, we extend computational creativity research to a culturally significant yet technologically under-served domain.

4. Methodology

This section details the corpus compilation, annotation procedure, evaluation metrics, and the experimental pipeline followed to analyze the generation quality of the sixth verse in *cordel sextilhas* using large language models (LLMs).

Initially, 18 *cordel* booklets were selected from two main online and public domain repositories[2] [3], and from these, a corpus comprising 1,519 *sextilhas* was created. These *sextilhas* served as the foundational data for subsequent annotation and experimentation. Each *sextilha* originally contained six verses, but for the purpose of this study, as we can see in Figure 1, the sixth (final) verse was removed and replaced by verses automatically generated by two LLMs: GPT-4o [OpenAI 2023] and LLaMA 3.2 (with 8 billion parameters) [Touvron et al. 2023].

```
{
  "id": "S0003",
  "verses": [
   "Eu recebi a mensagem",
   "Enviada por Mainha",
   "E garanto aos meus leitores",
   "Que não é invenção minha",
   "Porque eu sou um poeta"
  ]
 }
```

Figure 1. JSON example with the sixth (final) verse removed

To ensure reliability and consistency in evaluation, two human annotators independently assessed an initial subset of 150 *sextilhas*, annotating each generated verse

[2]https://www.netmundi.org/home/reliquias-do-cordel-38-obras-para-baixar/

[3]https://ler.ecordel.com.br

based on four evaluation dimensions: poetic meter, rhyme, coherence, and stylistic authenticity ("cordelisticity"). Each dimension was scored on a discrete scale with three possible values:

1. **Poetic Meter**:
   - 3: Perfect poetic meter (exactly 7 syllables).
   - 2: Acceptable rhythm (6 or 8 syllables).
   - 1: Poor rhythm (fewer than 6 or more than 8 syllables).
2. **Rhyme**:
   - 3: Perfect rhyme (matching the original stanza rhyme).
   - 2: Approximate or imperfect rhyme.
   - 1: No rhyme.
3. **Coherence**:
   - 3: Verse perfectly fits the theme and context.
   - 2: Verse is related but somewhat forced.
   - 1: Verse is incoherent or off-topic.
4. **Cordelisticity (Stylistic Authenticity)**:
   - 3: Strongly uses typical expressions, imagery, and popular language of *cordel* literature.
   - 2: Neutral language, lacking distinctive *cordel* traits.
   - 1: Excessively formal or artificial language.

The inter-annotator agreement, measured via Cohen's Kappa [Carletta 1996] (linear weighting), was exceptionally high, at 0.91, indicating strong confidence in annotation consistency. Given this high reliability, the annotators proceeded to expand annotations from the initial set of 150 *sextilhas* to the complete set of 1,519 *sextilhas*, ensuring a consistent evaluation coverage for subsequent analyses.

The corpus underwent several preprocessing steps. Initially, the *sextilhas* were extracted from plain-text files, manually reviewed, and standardized in terms of textual formatting. The *sextilhas* were then shuffled randomly to minimize any potential bias arising from the original ordering, ensuring that subsequent model training and evaluation would not be affected by text sequence effects. Following this, JSON batches were generated, each containing sets of 30 *sextilhas* formatted for efficient input to the LLMs. GPT-4o and LLaMA 3.2 models were then queried with structured prompts specifically designed to guide verse generation according to rhyme scheme, syllabic structure, coherence, and *cordel* stylistic authenticity.

In the preliminary analysis (150 annotated *sextilhas*), the rhyme accuracy metric (RhymeAcc, based on the agreement with human annotators) achieved approximately 72% (0.718). After annotation expansion to the complete corpus of 1,519 *sextilhas*, the rhyme accuracy further improved, reaching around 80% (0.793), with this improvement showing a better stability in model-generated rhyme quality across a larger and more representative corpus.

Regarding rhythm (poetic meter), an initial heuristic-based evaluation, relying on straightforward syllable counting via grapheme-to-phoneme conversion [Mortensen et al. 2018], revealed relatively low concordance with human annotators (around 11%). Consequently, several refinements were implemented to better match

human-perceived poetic rhythm. These adjustments included: (i) handling *sinalefa* (fusion of vowels between consecutive words), (ii) treating nasal diphthongs (such as "ão", "ãe") as single syllabic nuclei, (iii) adding a syllable adjustment for oxytone endings, and (iv) specific corrections for enclitic pronouns ("-me", "-te", "-lhe"). After these refinements, the rhythm accuracy (RhythmAcc) improved significantly, reaching approximately 71%, with a Cohen's Kappa of approximately 0.48 (moderate agreement) compared to human judgwments.

Finally, for model training and validation for RhythmAcc, the complete annotated corpus of 1,519 *sextilhas* was split into three sets: 1,063 *sextilhas* (70%) for training, 228 (15%) for development (validation), and 228 (15%) for the final hold-out test set. These splits were performed in a stratified manner, preserving the distribution of evaluation categories across subsets, with the subset of training and validation being merged and resplitted within a 5-fold cross-validation. The training and evaluation phases employed Logistic Regression classifiers trained on character n-gram representations and prosodic features (syllable counts, length, and stress patterns). The supervised classifier demonstrated a macro-averaged F1-score of approximately 0.747 and a Cohen's Kappa of about 0.66, clearly surpassing earlier heuristic methods.

5. Results and Discussion

This section presents the outcomes of our experiments, highlighting key results from both automatic evaluations and human annotations. Initially, the subset of 150 *sextilhas* evaluated by two expert annotators showed a substantial agreement (Cohen's Kappa linear-weighted at 0.91), reflecting strong reliability and consistency across annotators.

The initial automated evaluation indicated that the rhyme accuracy (RhymeAcc), measured by comparing LLM-generated sixth verses against original stanza rhymes, was approximately 71%. After expanding the analysis to the complete corpus of 1,519 *sextilhas*, the rhyme accuracy improved notably, reaching around 80%, suggesting that model-generated rhymes became more consistently accurate when trained and evaluated on a larger, more varied dataset.

Regarding poetic meter (rhythm), initial heuristic-based evaluations – using straightforward syllabic counting – yielded relatively poor concordance with human annotators (11%), leading us toto introduce several linguistically informed adjustments that enhanced the rhythmic evaluation, elevating RhythmAcc from an initial baseline of around 11% to approximately 71% post-adjustment, and achieving a moderate Cohen's Kappa of approximately 0.42 with human annotations.

Our supervised classification experiments utilized Logistic Regression models trained on the final annotated corpus, consisting of 1,063 training *sextilhas*, 228 for validation, and 228 for testing, with careful stratification. The classifier, leveraging character-level n-grams and prosodic features (syllable counts, stress positions, and verse length), achieved robust performance metrics: a macro-averaged F1-score of approximately 0.747 and a linear-weighted Cohen's Kappa of approximately 0.66, pointing to the effectiveness of machine learning approaches in capturing subtleties of poetic structure within *cordel* literature.

The coherence and stylistic authenticity (*cordelisticity*) metrics also displayed promising outcomes. While not the primary focus, annotators consistently rated coher-

ence highly, indicating that contemporary LLMs (GPT-4o [OpenAI 2023], LLaMA 3.2 [Touvron et al. 2023]) effectively maintained thematic consistency across generated *sextilha* verses. Stylistic authenticity varied more substantially, reflecting inherent challenges in capturing distinctive regional linguistic nuances of *cordel* through current generation models. Despite this, the overall ratings indicate a satisfactory level of stylistic adequacy for practical applications.

By releasing a richly annotated corpus and fully reproducible baselines, we deliver the first open platform that enables research on automatic verse generation – above all, for the Brazilian *cordel* tradition.

6. Final Remarks and Future Work

This work systematically explored the generation and evaluation of *cordel sextilha* verses using modern LLMs, implementing comprehensive annotations and linguistically informed evaluation metrics. Our experiments demonstrated significant progress in automating rhythm and rhyme evaluation, successfully achieving moderate-to-substantial agreement with expert human annotators. The final corpus of 1,519 annotated *sextilhas* constitutes an essential contribution, enabling further computational linguistics research focused on Brazilian *cordel* poetry.

Nonetheless, several challenges remain. Despite improved performance, rhythmic evaluation still showed limitations, primarily due to the inherent flexibility of poetic licenses (e.g. *sinalefa*, dialectal variations) and the subjective nature of poetic perception. Further improvements can be achieved through integrating more advanced grapheme-to-phoneme conversion models or by exploring neural network-based approaches for rhythm assessment.

Future studies may also benefit from applying Reinforcement Learning from Human Feedback (RLHF) methods, employing our established metrics as rewards to refine verse generation more interactively. Another promising direction involves developing and training larger-scale LLMs specifically fine-tuned on *cordel* corpora, potentially enhancing stylistic authenticity and regional linguistic nuance in verse generation.

Finally, considering our promising initial results, we intend to expand the annotation process, involving larger numbers *cordel* texts and diverse literary styles beyond the *sextilha* structure, ultimately fostering deeper computational engagement with Brazilian literary traditions.

References

Al-Rashid, F. and Ali, A. M. (2025). Automatic detection of quantitative metres in classical arabic poetry using bi-lstms. *Frontiers in Digital Humanities*, 2:1–12.

Araujo, A. P. R., Carlos, C. A. S. L., Sampaio, J. C. R., and Vieira, R. F. (2019). Digital heritage: Academic research in brazil in the last five years. In *Proceedings of the 27th CIPA International Symposium "Documenting the Past for a Better Future"*, volume XLII-2/W15, pages 109–116.

Campos, M. B., Tommaselli, A. M. G., Ivánová, I., and Billen, R. (2015). Data product specification proposal for architectural heritage documentation with photogrammetric techniques: A case study in brazil. *Remote Sensing*, 7(10):13337–13363.

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Chen, J., Patel, R., and Lee, M. (2024). Steering large language models toward poetic constraints via reinforcement learning from human feedback. *arXiv preprint*.

Gonçalo Oliveira, H., Cardoso, A., and Pereira, F. C. (2007). Tra-la-lyrics: An approach to generate text based on rhythm. In *Proceedings of the 4th International Joint Workshop on Computational Creativity*, pages 147–155, London, UK.

Greene, E., Bodrumlu, T., and Knight, K. (2010). Automatic analysis of rhythmic poetry with applications to generation and translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 524–533.

Instituto do Patrimônio Histórico e Artístico Nacional (IPHAN) (2018). Literatura de cordel becomes brazilian intangible cultural heritage. `https://portal.iphan.gov.br/noticias/detalhes/4833`. Accessed 23 Jun 2025.

Koziev, I. (2025). Automated evaluation of meter and rhyme in russian generative and human-authored poetry. *arXiv preprint*.

Martins, D. L., Lemos, D. L. d. S., Oliveira, L. F. R., Siqueira, J., do Carmo, D., and Medeiros, V. N. (2023). Information organization and representation in digital cultural heritage in brazil: Systematic mapping of information infrastructure in digital collections for data science applications. *Journal of the Association for Information Science and Technology*, 74(6):707–726.

Mittmann, A., von Wangenheim, A., and dos Santos, A. L. (2018). Aoidos: A system for the automatic scansion of poetry written in portuguese. In *Proceedings of the 19th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2018)*, volume 10762 of *Lecture Notes in Computer Science*, pages 611–628. Springer.

Mortensen, D. R., Dalmia, S., and Littell, P. (2018). Epitran: Precision g2p for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2710–2714. European Language Resources Association (ELRA).

Nogueira, A. G. R. (2018). Patrimonialização da literatura de cordel e os desafios de salvaguarda. *Anos 90*, 25(48):181–212.

OpenAI (2023). Gpt-4o: Technical report. `https://openai.com/research/gpt-4o`. Accessed: June 2025.

Pinheiro, F. F. A. (2023). O cordel contemporâneo: estudo do verso e poética. Dissertação de mestrado, Universidade Federal de São Paulo, São Paulo.

Plecháč, P., Cinková, S., Kolář, R., Šeła, A., Sisto, M. D., Nugues, L., Haider, T., and Kočnik, N. (2024). Poetree: Poetry treebanks in czech, english, french, german, hungarian, italian, portuguese, russian, slovenian and spanish. *Research Data Journal for the Humanities and Social Sciences*, 9:1–17.

Rigonatto, M. (2025). Redondilha: Concept and examples. `https://brasilescola.uol.com.br/o-que-e/portugues/o-que-e-redondilha.htm`. Accessed 23 Jun 2025.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Walsh, M., Preus, A., and Antoniak, M. (2024). Sonnet or not, bot? evaluating large language models on poetic form. *arXiv preprint*.

Wang, Z., You, K., Chen, J., and Zhao, S. (2016). Chinese poetry generation with planning-based neural network. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*, pages 1051–1060.

Yakovenko, N. (2020). Rhythm and rhyme evaluation toolkit for rap lyrics. `https://research.facebook.com/publications/rhyme-eval`. Accessed 23 Jun 2025.