Who Can Scan a Verse? Evaluating Open and Closed Language Models for Portuguese Poetic Scansion

Filipe Calegario, André Valença
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brazil
{fcac, aaav}@cin.ufpe.br

Abstract

Automatic poetic scansion, segmenting verses into metrical syllables and identifying stress patterns, is a key component for meter-aware computational poetry systems, yet it remains particularly challenging for Portuguese due to complex phonological phenomena. Previous work fine-tuned GPT-3.5 on this task (88.6% syllable ratio) but was limited to a single proprietary model. This paper presents the first systematic evaluation of 24 language models for Portuguese scansion in terms of accuracy, latency, and cost, with an additional pilot study of 9 reasoning models. We benchmark proprietary, open-weight, and Portuguese-specialized models in zero-shot and few-shot settings, and compare them against fine-tuned GPT-4.1 variants. We also introduce a stricter exact-match metric and format-tolerant metrics that together reveal substantially more phonological competence than prior evaluation suggested. Our key findings are: (1) fine-tuning remains essential, with GPT-4.1 variants surpassing the GPT-3.5 baseline on the same test set, and diminishing returns above a minimum model capacity; (2) in a pilot study, all nine reasoning models achieve 0% exact match, failing through distinct mechanisms (output truncation, empty responses, incorrect notation); and (3) Portuguese-specialized models underperform general-purpose alternatives, a gap that reflects weaker instruction-following more than weaker phonological competence. We discuss implications for computational creativity and accessibility pathways for communities that cannot depend on centralized commercial infrastructure.

Introduction

Poetic scansion, the division of verse into metrical syllables and the identification of stress patterns, is a foundational task for computational poetry systems (Oliveira 2017). It underpins meter classification, rhyme detection, and poetry generation, yet remains particularly challenging for Portuguese due to the language's rich phonological phenomena: synalepha, elision, diphthong reduction, and other metaplasms that alter syllable boundaries between the written and spoken forms of verse (Bisol 2005; Ali 1999).

Recent work has shown that a fine-tuned GPT-3.5-turbo can reach 88.6% syllable-level accuracy on Portuguese scansion when evaluated against the Aoidos rule-based corpus (Valença and Calegario 2025). However, that result was limited to a single model family and did not explore the rapidly expanding landscape of open-weight, multilingual, and Portuguese-specialized alternatives.

This paper presents a systematic evaluation of 24 language models for Portuguese poetic scansion across three axes: *accuracy*, *latency*, and *cost*, complemented by a pilot study of 9 reasoning models. We benchmark proprietary, open-weight, and Portuguese-specialized models in zero-shot and few-shot settings, and compare them against fine-tuned GPT-4.1 variants. Beyond the original syllable-ratio metric, we introduce both a stricter exact-match measure (SYL_EM%) and format-tolerant evaluation measures, including syllable boundary F1 (BND_F1), stress accuracy (STR_ACC), and metaplasm detection rate (META%), enabling finer-grained comparison even when models produce outputs that differ in surface formatting.

This work is part of a broader effort to build reliable analysis components that can underpin meter-aware poetry generation: a system that can accurately scan a verse is a prerequisite for one that can compose metrically valid verses. This is especially relevant for traditions such as Northeastern Brazilian popular poetry, where strict metric and rhyme conventions govern improvised oral performance.

Our contributions are threefold. (1) We present the first comparative benchmark of language models for Portuguese scansion (24 models across accuracy, latency, and cost, plus a pilot of 9 reasoning models), and introduce a stricter exact-match metric (SYL_EM%) and format-tolerant metrics (BND_F1, STR_ACC, META%) that reveal substantially more phonological competence than the syllable-ratio metric used in prior work. This reusable evaluation framework establishes that fine-tuning remains essential (the best fine-tuned model reaches 93.9% syllable ratio versus 74.1% for the best zero-shot model), with GPT-4.1 variants confirmed as a genuine improvement over the re-evaluated GPT-3.5 baseline. (2) We provide preliminary empirical evidence that extended reasoning can be counterproductive for format-sensitive linguistic tasks: in a pilot study (N=10), all 9 reasoning models reach 0% exact match, and we characterize three distinct failure mechanisms (output truncation, empty or leaked responses, and non-standard notation) that inform whether and how reasoning models should be de-

ployed in computational poetry pipelines. (3) We map the cost-accuracy trade-off across all model categories and find that Portuguese-specialized models underperform general-purpose alternatives, surfacing accessibility and linguistic-sovereignty concerns for communities that cannot depend on centralized commercial infrastructure.

Related Work

This section reviews prior work on automatic poetic scansion across languages, the use of large language models for poetry tasks, and the landscape of language models for Portuguese. Together, these strands reveal a significant gap: while computational scansion tools and LLM-based poetry evaluations have advanced considerably for several languages, Portuguese remains notably underserved.

Computational Scansion

Rule-based approaches have a long history in computational scansion, though unevenly distributed across languages. For English, early work dates back to Logan (1988) and Hayward (1996), who applied statistical and connectionist models to metrical analysis. Later systems such as Scandroid (Hartman 2005) and ZeuScansion (Agirrezabal et al. 2016), which combines finite-state technology with phonological dictionaries, achieved 86.78% per-syllable accuracy. Greene et al. (2010) used weighted finite-state transducers for rhythmic analysis with applications to poetry generation and translation.

For Spanish, Gervás (2000) proposed the first rule-based scansion system. Navarro-Colorado (Navarro-Colorado 2017) developed a fixed-metre scansion system and created the ADSO corpus of annotated Golden Age sonnets (Navarro-Colorado, Ribes Lafoz, and Sánchez 2016) with 96% inter-annotator agreement, enabling systematic benchmarking. De la Rosa et al. (2020) developed Rantanplan, the current state-of-the-art, achieving 96.23% accuracy on hendecasyllabic verse. Marco Remón and Gonzalo (2020) later demonstrated that scansion without syllabification can outperform traditional pipelines while running over 20 times faster.

For Portuguese, the landscape is notably sparse. Araújo and Mamede (2002) presented an early poem classifier in 2002, but the only comprehensive scansion system is Aoidos (Mittmann 2016; Mittmann and Maia 2017), which applies phonological rules to segment verses into poetic syllables and identify stress patterns. Validated on over 100,000 verses spanning three centuries, Aoidos provides the rule-based scansions that serve as ground truth. Building on this resource, our own prior work (Valença and Calegario 2025) organized these scansions into a structured dataset with stress annotations and fine-tuned GPT-3.5-turbo on it, achieving 88.6% syllable ratio and establishing the baseline and evaluation framework that the present work extends.

This asymmetry is striking: Spanish has at least three dedicated scansion systems, English has several, and even Basque (Agirrezabal et al. 2012) and German (Bobenhausen 2011) have specialized tools. The most comprehensive cross-lingual scansion studies (Agirrezabal's multilingual thesis (2017) and De la Rosa et al.'s transformer experiments across Spanish, English, and German (2023)) did not include Portuguese, despite it being spoken by over 250 million people.

LLMs for Linguistic and Poetic Tasks

Large language models have demonstrated remarkable few-shot capabilities across linguistic tasks (Brown and others 2020), and chain-of-thought prompting has been shown to improve performance on reasoning-intensive problems (Wei et al. 2022). However, phonologically demanding tasks like scansion require precise character-level manipulation that challenges autoregressive generation, especially for languages underrepresented in training corpora. Belouadi and Eger (2023) showed that token-free character-level models can outperform token-based LLMs on poetry tasks, precisely because token boundaries do not align with syllable or phoneme boundaries. Similar evidence comes from Chinese classical poetry, where, for the related task of format-controlled generation, token-free architectures substantially outperform token-based LLMs such as GPT-4 (Yu et al. 2024), underscoring tokenization as a cross-lingual bottleneck for strictly metered tasks.

Recent evaluation work has begun to assess LLM poetic capabilities systematically. Walsh et al. (2024) evaluated LLM recognition of over 20 English poetic forms. Koziev (2025) introduced a scansion tool and evaluation dataset for Russian poetry. Chatzikyriakidis and Natsina (2026) found that pure LLM generation for Greek poetry fails catastrophically (under 4% valid poems), but hybrid systems combining LLMs with phonological algorithms restore performance to 73.1%. This "Reasoning Gap" suggests that LLMs require specialized support for phonological tasks in non-English languages, a finding directly relevant to Portuguese scansion.

Poetry, Computational Creativity, and Accessibility

Computational poetry generation has explored meter-aware systems (Colton, Goodwin, and Veale 2012; Lau et al. 2018; Oliveira 2012; Ormazabal et al. 2022), and instruction-tuned models have been applied to collaborative poetry writing (Chakrabarty, Padmakumar, and He 2022). Gonçalo Oliveira (2024), in a recent survey of computational creativity systems for Portuguese, highlighted both the progress in platforms like PoeTryMe and the persistent scarcity of tools for this language. However, accurate scansion remains a prerequisite for meter-constrained generation, automatic verse evaluation, and human-AI co-creation of formal poetry. Without reliable automated scansion, such systems must rely on hand-crafted phonological rules or human evaluation, limiting their scalability and applicability to under-resourced languages like Portuguese.

The accessibility of these tools also matters. While the Sabiá series (Pires et al. 2023; Almeida et al. 2024; Abonizio et al. 2025) demonstrated cost-effective fine-tuning for Portuguese, these models remain behind commercial APIs. Open alternatives such as Tucano (2025) (up to 2.4B parameters, Apache 2.0) and TeenyTinyLlama (Corrêa et al. 2024) (up to 460M parameters, developed for under

$500) show that accessible Portuguese language models are emerging. These models were not included in our benchmark because they are base (non-instruction-tuned) models too small to follow zero-shot or few-shot prompts; however, they represent promising candidates for future fine-tuning experiments. The present study provides empirical evidence on how instruction-following models compare on a concrete, phonologically demanding task, informing decisions about building creativity support tools that do not depend on centralized commercial infrastructure.

Methodology

This section defines the scansion task, describes the dataset and models under evaluation, and introduces the metrics used to assess performance.

Task Definition

Given a Portuguese verse as input, the scansion task requires producing a string that segments the verse into *poetic* syllables separated by "/", marks stressed syllables with "⋆", and indicates the last metrically relevant stressed syllable with "#". For example:

Input	*Teus formosos quinze anos*
Output	`Teus / for- / ⋆mo- / sos / quin- / ⋆ze a-# / nos`

Poetic syllables differ from grammatical ones due to phonological phenomena called *metaplasms*. The verse above has 8 grammatical syllables (*Teus / for / mo / sos / quin / ze / a / nos*), but only 7 poetic syllables (6 metric + 1 post-tonic). The reduction occurs because the final vowel of *quinze* and the initial vowel of *anos* merge through *synalepha*, producing the fused syllable "`ze a-`". Stress markers identify the tonic syllables: "`⋆mo-`" (from *for**mo**sos*, position 3) and "`⋆ze a-`" (carrying the stress of ***a**nos*, position 6). The "#" on position 6 signals the metric boundary: Portuguese verse counts syllables only up to the last stressed one, making this a hexasyllable (6-syllable verse).

The notation thus captures three distinct linguistic phenomena: syllable segmentation (boundaries), stress assignment (tonic identification), and metaplasm detection (cross-word syllable fusion). Metaplasms are the most challenging aspect, as they require phonological knowledge to determine when adjacent vowels across word boundaries merge into a single poetic syllable (Bisol 2005; Ali 1999).

This task formulation targets stress-and-syllable-based prosody, characteristic of Romance and Germanic verse traditions, and is not intended to generalize to all poetic systems. Classical Arabic metrics (*ʿarūḍ*), for instance, are built on moraic units defined by vowel quantity rather than on lexical stress (Frolov 2000). Our notation, metrics, and conclusions are therefore language-specific by design.

Dataset

Our dataset derives from Aoidos (Mittmann 2016), a rule-based phonological system for Portuguese scansion available as an online platform.

In our prior work (Valença and Calegario 2025), we constructed the dataset by taking approximately 13,500 verses from the Aoidos test corpus (spanning three centuries of Brazilian Portuguese poetry, publicly available in the Aoidos repository), submitting them in batches to the Aoidos online platform, and parsing the resulting HTML output with a custom script to extract syllable segmentations, metrical classifications, and rhythmic patterns.

Stress markers (⋆ and #) were then added programmatically: for each verse, Aoidos reports the rhythmic pattern as the positions of the metrically stressed syllables, and a script maps these positions onto the segmented syllable sequence, prepending ⋆ to every stressed syllable and appending # to the last one to mark the metric boundary. The annotation script is provided in the supplementary material. Verses containing diérese were filtered out to ensure notational consistency, and the results were deduplicated, yielding 14,080 unique verses with ground-truth scansions.

Following the same split as our prior work, the dataset is divided 75/25 into 10,560 training and 3,520 test verses. Fine-tuned models are trained on 7,040 examples drawn from the training split (the same subset used in our prior work). From the 3,520 test verses, we sample 500 (random seed 42) as a canonical test set, balancing evaluation cost with statistical power.

All models are evaluated on the same 500 verses to ensure comparability. To prevent data leakage, we verified that the 500 test verses have zero overlap with the fine-tuning training and validation sets.

Models

Model selection was guided by API availability and practical accessibility: all models were accessed through cloud APIs (OpenAI, Anthropic, Google, DeepSeek, Maritaca AI, and OpenRouter for open-weight models), and no models were run locally. This reflects a deliberate design choice to evaluate the landscape as it is available to most researchers and practitioners; local deployment of open-weight models remains an avenue for future work. We evaluate 24 models in the main benchmark, organized into four categories.

Fine-tuned (3 models): GPT-4.1, GPT-4.1-mini, and GPT-4.1-nano, fine-tuned on the 7,040 training examples described above, using the same system prompt and output format as our prior work (Valença and Calegario 2025).

Proprietary (9 models): Claude Opus 4.6, Sonnet 4.6, and Haiku 4.5 (Anthropic 2026); GPT-4.1, GPT-4.1-mini, and GPT-4.1-nano (OpenAI 2023); DeepSeek V3 (DeepSeek-AI 2024); Gemini 2.5 Flash and Flash Lite.

Open-weight (7 models): LLaMA 4 Maverick and Scout (Meta AI 2025); Gemma 3 12B and 4B (Google DeepMind 2024); Ministral 14B, 8B, and 3B (Mistral AI 2025).

Portuguese-specialized (5 models): Sabiá 4, 3.1, and 3; Sabiazinho 4 and 3 (Pires et al. 2023; Abonizio et al. 2025).

Additionally, 9 reasoning/thinking models were evaluated on a 10-example pilot (see *Thinking Models* in Results): DeepSeek-R1, Gemini 2.5 Pro, GPT-5 Nano, Grok-4 Fast, Kimi-K2.5, Qwen3.5-27B, Qwen3-32B, Qwen3-14B, and Qwen3-8B. GPT-5 full and GPT-5-mini were tested in preliminary experiments but excluded due to prohibitive latency

(∼98s/verse) and cost ($0.042/verse for GPT-5 full); GPT-5 Nano was retained as a representative of the family at acceptable cost.

Model Settings and Prompting All evaluations use temperature 0.1, following our prior work (Valença and Calegario 2025). A near-zero temperature favors deterministic outputs while avoiding the degenerate responses that temperature 0 elicits from some providers; we quantify the resulting run-to-run consistency in the appendix.

The maximum output length is set to 512 tokens for standard models. For thinking/reasoning models, the token budget was increased to 8,192 completion tokens to accommodate the internal chain-of-thought (thinking mode), and the temperature parameter was omitted as these APIs do not support it. Note that, in some cases, extended thinking is a configurable mode rather than an inherent model property: Claude Opus 4.6, for instance, supports extended thinking but was evaluated with thinking disabled and is therefore classified as a standard proprietary model in this study.

The zero-shot system prompt was carefully revised to disambiguate formatting conventions (e.g., explicit rules for spacing around stress markers and hyphens), as preliminary experiments revealed that the original prompt's ambiguity unfairly penalized zero-shot models on exact match. The prompt specifies the output format but deliberately omits phonological rules, so that the evaluation measures the model's linguistic competence rather than its ability to follow detailed instructions.

Few-shot prompts include 3 worked examples that demonstrate the input/output format across diverse metaplasm types. Fine-tuned models use the original system prompt from our prior work. Throughout the paper we abbreviate these three settings as **ZS** (zero-shot), **FS** (few-shot), and **FT** (fine-tuned). To conserve space, Table 1 reports for each model only its best-performing configuration: fine-tuned models are labeled FT, while every non-fine-tuned model reached its best result in the few-shot setting and is therefore labeled FS. The corresponding zero-shot scores can be recovered by subtracting the few-shot gains reported in Table 2(a), and complete per-mode results are provided in the supplementary material. Open-weight models are accessed via OpenRouter; all other models use their respective provider APIs. All experiments were conducted between December 2025 and February 2026. Evaluation scripts, prompts, raw model outputs, and the test set are available as supplementary material (`https://github.com/filipecalegario/portuguese-poetic-scansion-benchmark`).

Evaluation Metrics

We employ seven metrics spanning four dimensions:

Syllable accuracy: We report two complementary metrics. *SYL_R%* (syllable ratio), from prior work (Valença and Calegario 2025), compares syllables position-by-position up to the last stressed syllable and computes the average proportion of matching syllables per verse. *SYL_EM%* (exact match) is a stricter metric introduced in this work: a verse scores 1 only if every syllable, stress marker, and formatting detail matches the ground truth exactly. The difference between these metrics is informative: a model scoring 88% SYL_R% but 62% SYL_EM% produces mostly correct syllables in each verse but rarely achieves a perfect transcription. SYL_R% enables direct comparison with the baseline from prior work; SYL_EM% indicates the proportion of scansion that is production-ready without post-processing.

Boundary quality: *BND_F1%*, the F1 score computed by aligning predicted and expected syllable sequences via edit distance; true positives are syllables with matching normalized text. Unlike SYL_R%, which uses positional alignment (via `zip`), BND_F1 uses edit-distance alignment and thus handles insertions and deletions more gracefully, providing a fairer comparison between fine-tuned models (which learn the exact corpus format) and zero-shot models (which may produce different syllable counts).

Stress quality: *STR_ACC%*, the percentage of correctly aligned syllables where the stress marker matches. When prediction and reference differ in length, syllables are aligned by the same edit-distance procedure used for BND_F1, and stress is compared only on pairs whose normalized text matches, so that stress accuracy is not confounded by segmentation errors. *STR_PAT%*, a binary stress pattern match up to the last stressed syllable.

Metaplasm detection: *META%*. We operationalize a metaplasm as a poetic syllable that spans two or more words through cross-word vowel fusion, which in our notation surfaces as an internal whitespace inside a syllable token (e.g., `que en` for synalepha, or `se há` for elision). For a verse, let M be the set of such fused syllables in the ground truth. A ground-truth metaplasm counts as recovered if some predicted syllable has the same normalized text (lowercased, with stress markers, trailing hyphens, and punctuation removed), regardless of its position in the line. META% is the number of recovered metaplasms divided by $\sum |M|$, aggregated over all verses that contain at least one metaplasm (verses with none are excluded). This operationalization locates cross-word fusion, the most linguistically challenging aspect of Portuguese scansion, but it does not classify the underlying phonological process (synalepha, elision, crasis) and requires the fused syllable to match the reference text exactly.

Efficiency: *Latency* (wall-clock seconds per API call, averaged over all 500 verses) and *Cost* (USD per verse). Cost is estimated as $(T_{\text{in}} \times p_{\text{in}} + T_{\text{out}} \times p_{\text{out}})/N$, where T_{in} and T_{out} are the total input and output tokens reported by the API, p_{in} and p_{out} are the provider's published per-token prices, and $N = 500$ is the number of verses. Input tokens include the system prompt, any few-shot examples, and the verse itself; output tokens cover the model's scansion response.

Results

Table 1 presents the main results across all model categories, showing the best configuration (zero-shot or few-shot) for each model. Table 2 details few-shot gains and thinking model performance. Figure 1 visualizes the accuracy-cost trade-off, and Figure 2 compares the best score per category across key metrics.

Table 1: Main results across model categories. Best mode (zero-shot or few-shot) per model. SYL_R% is the syllable-level ratio from prior work; SYL_EM% is our stricter verse-level exact match. Best value per column in **bold**.

Model	Mode	SYL_R↑	SYL_EM↑	BND_F1↑	META↑	STR_ACC↑	Lat(s)↓	$/verse↓
Fine-Tuned								
GPT-4.1-mini-FT	FT	**93.9**	81.8	**98.0**	**96.6**	98.0	2.2	0.00027
GPT-4.1-FT	FT	93.3	**82.4**	97.9	96.2	**98.4**	3.9	0.00102
GPT-3.5-FT	FT	88.0	62.2	95.7	92.5	96.0	1.7	0.00085
GPT-4.1-nano-FT	FT	76.1	34.8	90.4	82.0	91.4	0.8	0.00007
Proprietary								
Claude Opus 4.6	FS	74.1	25.8	86.8	67.6	94.4	2.8	0.00312
Claude Sonnet 4.6	FS	63.9	17.6	83.0	62.8	90.7	2.1	0.00187
Gemini 2.5 Flash	FS	63.9	15.4	71.9	47.5	90.5	1.8	0.00007
GPT-4.1	FS	55.0	9.2	76.5	53.9	89.0	1.2	0.00102
Gemini 2.5 Flash Lite	FS	52.8	6.4	80.3	49.5	84.1	0.7	0.00005
Claude Haiku 4.5	FS	51.2	7.4	79.6	48.2	81.6	1.0	0.00062
DeepSeek V3	FS	46.5	7.6	76.7	26.5	86.4	2.4	0.00012
GPT-4.1-mini	FS	40.6	3.6	72.9	27.9	79.0	1.2	0.00021
GPT-4.1-nano	FS	32.6	0.0	62.6	23.6	71.2	1.0	0.00005
Open-Weight								
LLaMA 4 Maverick	FS	45.9	7.4	74.8	34.7	85.5	1.3	0.00008
Ministral 14B	FS	40.0	2.6	73.1	29.9	78.7	0.8	0.00008
LLaMA 4 Scout	FS	37.5	1.8	70.8	28.9	75.7	1.2	0.00004
Gemma 3 12B	FS	33.0	0.8	68.6	22.6	72.3	1.4	0.00002
Ministral 8B	FS	30.7	1.4	65.9	13.5	70.2	0.6	0.00006
Gemma 3 4B	FS	23.7	0.0	50.4	16.2	66.8	2.4	**0.00002**
Ministral 3B	FS	22.4	0.0	52.3	21.2	68.0	**0.6**	0.00004
Portuguese-Specialized								
Sabiá 4	FS	45.8	3.4	74.8	43.3	75.9	1.5	0.00045
Sabiá 3	FS	33.5	0.8	63.7	36.7	68.8	2.6	0.00038
Sabiá 3.1	FS	31.5	0.6	66.2	44.3	65.8	2.3	0.00038
Sabiazinho 4	FS	30.5	0.8	63.8	31.9	67.1	0.7	0.00009
Sabiazinho 3	FS	26.6	1.0	55.6	25.8	64.4	1.1	0.00008

Fine-Tuning Dominates

Fine-tuned models dominate all metrics, as shown in the top group of Table 1. GPT-4.1-mini-FT leads on SYL_R% (93.9%), closely followed by GPT-4.1-FT (93.3%), both with BND_F1 $>$ 97% and META% $>$ 96%. The near-equivalence of these two models suggests that the scansion task may be saturated with fine-tuning and does not benefit from the additional capacity of larger architectures. However, GPT-4.1-nano-FT collapses to 76.1% SYL_R% (34.8% SYL_EM%), indicating a minimum model capacity threshold for this task.

A key finding is that re-evaluating the GPT-3.5-FT model from our prior work (Valença and Calegario 2025) on the same 500-verse test set yields 88.0% SYL_R%, reproducing the original 88.6% baseline and confirming that GPT-4.1 variants represent improvement (+5.3 percentage points on SYL_R%), not a regression. The difference between SYL_R% and SYL_EM% is also informative: GPT-3.5-FT scores 88.0% on the former but only 62.2% on the latter, meaning it produces mostly correct syllables but achieves a perfect transcription less often than GPT-4.1-FT (82.4% SYL_EM%).

As visible in Figure 2, the gap between fine-tuned and zero-shot models is stark across all metrics, with fine-tuned models achieving approximately 98% BND_F1 versus 87% for the best proprietary model.

Zero-Shot and Few-Shot Performance

Among non-fine-tuned models, Claude Opus 4.6 achieves the highest SYL_R% (74.1% few-shot) and SYL_EM% (25.8%), followed by Claude Sonnet 4.6 (63.9% / 17.6%) and Gemini 2.5 Flash (63.9% / 15.4%). While the exact-match scores are low, the syllable-ratio and BND_F1 metrics reveal substantially more competence: Claude Opus 4.6 reaches 86.8% BND_F1, meaning it correctly segments the majority of syllable boundaries but fails on stress placement or formatting details.

Table 2(a) quantifies the few-shot effect. All five selected models improve with 3-shot examples, with gains ranging from +3.0 to +9.2 percentage points on SYL_EM% and +0.7 to +13.4 on BND_F1. The largest relative gains are observed for META% (+7.3 to +23.3 points), indicating that few-shot examples can be particularly helpful for models to learn metaplasm conventions.

Figure 1 reveals the accuracy-cost trade-off landscape. The Pareto frontier traces from the cheapest models (Gemma 3 12B at $0.00002/verse with 68.6% BND_F1) through mid-range options (DeepSeek V3 at $0.0001 with

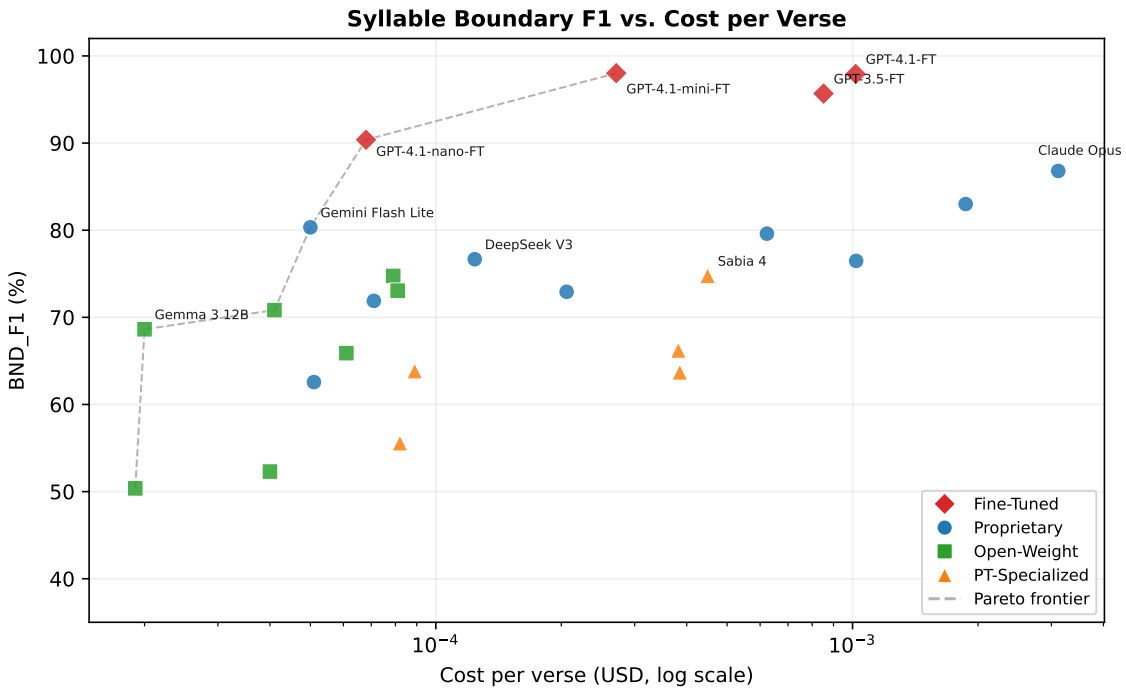

Figure 1: Syllable boundary F1 score vs. cost per verse (USD, log scale). Colors indicate model category: fine-tuned (red), proprietary (blue), open-weight (green), Portuguese-specialized (orange). The dashed line shows the Pareto frontier.

76.7%) to the premium tier (Claude Opus 4.6 at $0.003 with 86.8%). Fine-tuned models occupy the upper-right region, combining high accuracy with moderate cost.

Thinking Models: Preliminary Evidence of Failure

Table 2(b) reports results for 9 reasoning/thinking models evaluated on a 10-example pilot. The reduced sample size reflects the prohibitive cost of full evaluation: with latencies of 7–211s per verse (vs. <2s for standard models), a complete 500-verse evaluation across all 9 thinking models would require an estimated 40–200 hours of API time. Every model achieves 0% SYL_EM%, with costs up to $0.029/verse. Because exact match is unforgiving of partial or truncated outputs, we report it alongside two edit-distance-based measures, BND_F1 and normalized edit distance (NRM_ED, where 0 is identical), which together reveal that several models retain substantial syllabification content despite the 0% exact-match score.

Qualitative analysis of the outputs reveals that the 0% exact match stems from at least three distinct failure modes, not a single cause:

(1) Output truncation: Gemini 3.1 Pro allocates the vast majority of its token budget to internal reasoning (∼488 reasoning tokens) and produces only ∼20 completion tokens, cutting every scansion mid-output (e.g., `por / ⋆ vos- / sos / ⋆ fi- / lhos / cho- / ⋆ rar /`, missing the final syllables). The reasoning-to-completion ratio reaches 24.5×.

(2) Empty or leaked output: GPT-5 Nano produces empty responses for 80% of verses despite consuming ∼8,128 reasoning tokens per call (notably, with the default output limit of 512 tokens, 100% of responses were empty, as the reasoning phase alone exhausted the entire budget before any visible output could be generated). Similarly, Qwen 3.5 27B occasionally emits its internal chain-of-thought (e.g., "Thinking Process:") instead of the scansion, spending ∼14,323 reasoning tokens at 211s/verse without converging on an answer.

(3) Incorrect notation: Qwen3-32B and Qwen3-8B produce non-empty outputs but deviate from the required notation (e.g., placing stress markers after syllables instead of before: `vós ⋆` instead of `⋆ vós`, or merging syllables: `⋆vos-sos` instead of `⋆ vo- / ssos`).

Among the remaining models (DeepSeek-R1, Grok-4 Fast, Kimi-K2.5), the outputs follow the correct format, and the edit-distance metrics confirm that the failure is one of completion and formatting rather than phonology: DeepSeek-R1 reaches the highest BND_F1 (77.5%) at a normalized edit distance of only 0.25 (roughly three-quarters of the characters match the reference), with Grok-4 Fast and Kimi-K2.5 at NRM_ED 0.31 and 0.42. Errors in syllable boundaries and stress placement, not an inability to segment, are what prevent an exact match.

While the small sample size (N=10) limits generalizability, the pattern suggests that the extended reasoning process can be counterproductive for scansion: models spend thousands of tokens deliberating about phonological rules but fail to produce correctly formatted output, either by exhausting their token budget on reasoning, by not converging on a response, or by inventing non-standard notation conventions.

Table 2: (a) Few-shot gain on selected models. (b) Reasoning/thinking models (all N=10; all score 0% SYL_EM). NRM_ED is the normalized character-level edit distance to the ground truth (0 = identical; lower is better). Best per column in **bold**.

(a) Few-Shot Gain

Model	ΔR↑	ΔEM↑	ΔF1↑	ΔMETA↑
DeepSeek V3	**+13.6**	+6.8	+11.5	**+23.3**
LLaMA 4 Mav.	+8.8	+5.8	+7.0	+17.8
GPT-4.1	+4.5	+7.2	+9.6	+17.2
Cl. Opus 4.6	+4.1	**+9.2**	+0.7	+3.7
Sabiá 4	+1.7	+3.0	**+13.4**	+7.3

ΔR = ΔSYL_R, ΔEM = ΔSYL_EM, ΔF1 = ΔBND_F1

(b) Reasoning / Thinking Models

Model	SYL_R	BND_F1	NRM_ED↓	Lat	$/v
Gemini 3.1 Pro	**63.0**	59.6	0.53	**6.8**	0.0007
DeepSeek-R1	41.4	**77.5**	**0.25**	89	0.0015
Qwen 3.5 27B	35.4	66.6	0.35	184	0.0292
Grok 4 Fast	18.7	73.4	0.31	8.6	0.0006
Kimi K2.5	12.2	61.5	0.42	96	0.0118
Qwen 3 32B	11.7	45.7	0.59	87	0.0007
Qwen 3 14B	10.3	39.2	0.64	25	**0.0003**
Qwen 3 8B	9.9	34.0	0.70	54	0.0010
GPT-5 Nano	6.5	13.1	0.89	49	0.0033

Several of these failures are configuration-dependent rather than fundamental: raising the completion-token budget or adjusting the prompt to suppress chain-of-thought leakage would likely recover many exact matches, since the underlying syllabification (as reflected by NRM_ED and BND_F1) is often largely correct. We therefore present this pilot as preliminary evidence motivating a fuller study, not as a definitive verdict on reasoning models.

Portuguese-Specialized Models Underperform

Portuguese-specialized models (Sabiá family) do not outperform general-purpose alternatives on this task. The best Portuguese model, Sabiá 4, reaches 45.8% SYL_R% and 74.8% BND_F1 in few-shot mode. Notably, while its exact-match-oriented SYL_R% trails general-purpose models, its format-tolerant BND_F1 is comparable to several of them (e.g., GPT-4.1 at 76.5% and DeepSeek V3 at 76.7%, and above Gemini 2.5 Flash at 71.9%), indicating that the gap is concentrated in formatting and notation adherence rather than in syllable segmentation. This pattern is visible in both Table 1 (PT group vs. Proprietary group) and Figure 2.

Stress Is Easier Than Segmentation

An informative pattern emerges from the STR_ACC metric in Table 1: even models with low SYL_EM% achieve high stress accuracy on correctly segmented syllables. For example, Claude Haiku (7.4% SYL_EM%) reaches 81.6% STR_ACC, and DeepSeek V3 (7.6% SYL_EM%) reaches 86.4%. This indicates that models have implicit knowledge of Portuguese stress patterns but struggle with the more challenging syllable segmentation and metaplasm detection.

Table 3: Scansion outputs for the verse *"Verdade, que entre vós se há confundido,"*. Differences from the ground truth are underlined.

Source	Scansion
Ground truth	Ver- / *da- / de, / *que en- / tre / *vós / se há / con- / fun- / *di-# / do,
GPT-4.1-FT	Ver- / *da- / de, / *que en- / tre / *vós / se há / con- / fun- / *di-# / do,
Cl. Opus 4.6 (FS)	Ver- / *da- / de, / que en- / tre / *vós / se / *há / con- / fun- / *di-# / do,
Gemma 3 12B	Ver- / *da- / de, / que / en- / *tre / vós / se / *há / con- / *fun-# / di- / do
Sabiá 4	*Ver- / da- / de, / que en- / *tre / vós / se há / *con- / fun- / *di-# / do,

Qualitative Error Example

Table 3 illustrates how different models handle the verse *"Verdade, que entre vós se há confundido,"*, which contains two metaplasms: "que en-" (synalepha) and "se há" (elision). GPT-4.1-FT reproduces the ground truth exactly. Claude Opus 4.6 correctly detects "que en-" but splits "se há" into two separate syllables, a boundary error that costs the exact match. Gemma 3 12B misses both metaplasms and places the metric boundary marker (#) on the wrong syllable. Sabiá 4 detects the "se há" metaplasm but misplaces stress markers throughout. These examples illustrate the typical failure gradient: formatting and stress errors (Claude Opus 4.6) are less severe than boundary errors (Gemma, Sabiá), which the BND_F1 metric captures more faithfully than exact match.

Discussion

The results raise several questions about what drives scansion performance, what the metrics reveal beyond surface accuracy, and what the practical implications are for building accessible creativity support tools.

The Fine-Tuning Gap

Our results confirm that fine-tuning remains indispensable for production-quality Portuguese scansion. The gap between fine-tuned models (BND_F1 $\approx$ 98%) and the best zero-shot model (BND_F1 = 86.8%) is substantial and consistent across metrics, as illustrated in Figure 2. This aligns with the general finding that domain-specific tasks requiring precise formatting and linguistic conventions benefit disproportionately from supervised adaptation (Hu et al. 2022).

However, the near-equivalence of GPT-4.1-FT and GPT-4.1-mini-FT (differing by only 0.6 percentage points on SYL_R%) suggests that the task's complexity does not warrant the largest model architectures, a practically important finding for deployment cost. At $0.0003/verse, GPT-4.1-mini-FT offers the best cost-accuracy trade-off among fine-tuned models.

What BND_F1 Reveals

The introduction of format-tolerant metrics like BND_F1 substantially changes the picture painted by exact match

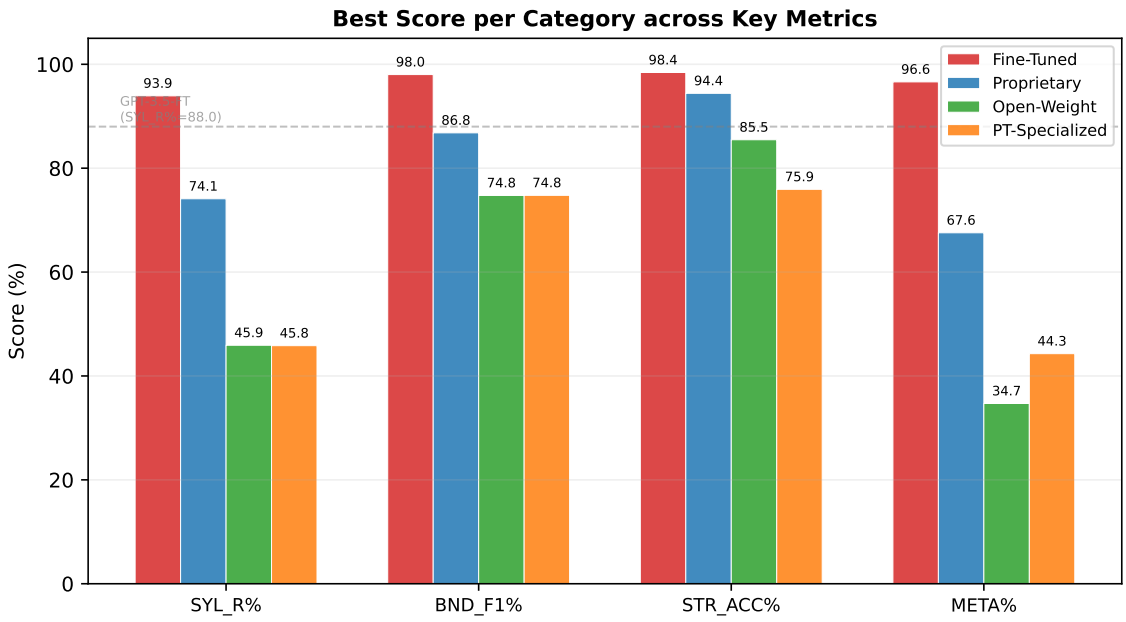

Figure 2: Best score per model category across key metrics. The dashed line marks the GPT-3.5 fine-tuned baseline (SYL_R% = 88.0%, measured on the same 500-verse test set).

alone. Under SYL_EM%, the best zero-shot model (Claude Opus 4.6 at 25.8%) appears to perform poorly. Under BND_F1, the same model scores 86.8%, revealing that it correctly identifies most syllable boundaries but fails on the precise formatting conventions (stress marker placement, hash notation) that the exact-match metric demands. This suggests that LLMs possess substantial implicit phonological knowledge of Portuguese but struggle to produce outputs that conform to a specific annotation standard without supervised examples.

This raises a natural question: does BND_F1 measure syllabification or merely formatting? By construction, it measures the former. Before alignment, both syllable sequences are lowercased and stripped of stress markers, hyphens, and punctuation, so two syllables match only when their underlying text is identical, irrespective of surface formatting. BND_F1 therefore credits correct *segmentation* of the Portuguese syllable string rather than adherence to the annotation convention, and a model gains nothing from English-style formatting habits, since the comparison is over Portuguese poetic syllables. The gap between SYL_EM% and BND_F1 isolates the formatting layer, while STR_ACC% and META% isolate stress assignment and metaplasm detection; read together, the metric suite separates genuine phonological competence from the ability to reproduce a specific annotation standard.

Implications for Computational Creativity

Our results have direct consequences for computational poetry systems. Meter-constrained generation (Lau et al. 2018; Ormazabal et al. 2022) requires a scansion module to verify whether candidate verses satisfy target meters (e.g., decasyllables). With the best fine-tuned model achieving 98% BND_F1, such a filter becomes practical: a generator can produce candidate verses and use the scansion model to accept or reject them based on metrical criteria. Even zero-shot models, with BND_F1 scores above 80%, could serve as approximate filters in interactive co-creation settings where a human poet refines the output.

This is particularly relevant to poetic traditions in which meter is not optional but constitutive. In Northeastern Brazilian popular poetry, *cantadores* (improvising poets) engage in *pelejas*, oral verse duels governed by strict metric and rhyme conventions. As a poet from this tradition once put it, there are three pillars a *cantador* must master: rhyme, meter, and subject matter (Caldas 2021). A computational system aspiring to participate in such traditions, whether as a creative partner, a training tool for aspiring poets, or an analytical aid for literary scholarship, must first demonstrate reliable metrical analysis. The scansion benchmark presented here thus serves as a foundation for our broader platform, Pajéu (Valença and Calegario 2025), aimed at meter-aware verse generation, literary games, and human-AI co-creation rooted in these living poetic traditions.

Accessibility and Linguistic Sovereignty

The cost-accuracy landscape in Figure 1 reveals a concerning pattern for accessibility. The most accurate models are either fine-tuned proprietary models (requiring API access to OpenAI) or expensive frontier models (Claude Opus 4.6 at $0.003/verse). Open-weight alternatives, while dramatically cheaper (Gemma 3 12B at $0.00002/verse), sacrifice significant accuracy.

More concerning is the linguistic sovereignty dimen-

sion, which is the non-dependence on foreign infrastructure. The highest-performing models are developed by organizations in the United States and China, whereas Portuguese-specialized models from Brazilian companies do not lead on this task. This gap, however, is less a deficit of Portuguese phonology than of instruction-following: the Sabiá-3 technical report itself notes that the model lags general-purpose alternatives, including cheaper ones, on instruction-following benchmarks such as IFeval (Abonizio et al. 2025), and our format-tolerant BND_F1 shows Sabiá's syllable segmentation to be comparable to general-purpose models. The underperformance is therefore concentrated in adhering to the scansion notation rather than in Portuguese phonology per se. The practical implication for sovereignty nonetheless stands: no accessible Portuguese-tuned model currently delivers production-quality scansion. A promising pathway is fine-tuning open-weight models on the Aoidos corpus, which could combine the accessibility of open models with the accuracy of supervised adaptation, though the infrastructure and expertise required for fine-tuning remain barriers for many poetry communities.

Limitations

Several limitations should be acknowledged. First, our ground truth is obtained from a rule-based system (Aoidos) rather than from human expert annotation. This means our evaluation measures agreement with Aoidos's phonological rules, not necessarily with all valid scansion interpretations (some verses admit multiple legitimate scansions depending on performance choices). Fine-tuned models may learn to replicate Aoidos's systematic biases. We adopt Aoidos as an *operational* ground truth, consistent with our prior work (Valença and Calegario 2025) and the original Aoidos validation (Mittmann 2016), while acknowledging that a human-validated gold set would strengthen future evaluations.

Second, thinking models were evaluated on only 10 examples due to prohibitive cost and latency; while the consistent 0% SYL_EM% across all nine models is an initial signal, a larger sample would strengthen this finding.

Third, we evaluated only one prompting strategy per mode; more elaborate prompt engineering (e.g., including explicit phonological rules) might improve zero-shot performance, particularly on formatting conventions.

Fourth, our cost estimates use published list pricing as of February 2026 and do not account for volume discounts or cached responses.

Fifth, training and test verses are drawn from the same corpus and thus share a metrical distribution dominated by decasyllabic meters (*heroico* and *sáfico*, together about 70% of verses) and heptasyllables (about 15%), spanning 20 classical metrical classes. Our results therefore measure in-distribution performance and do not test generalization to genuinely novel structures, such as haiku, *cordel* improvisation, or metrically free verse; evaluating fine-tuned models on such out-of-distribution forms is an important direction for future work.

Conclusion

We have presented the first comparative evaluation of language models for Portuguese poetic scansion, benchmarking 24 models across accuracy, latency, and cost, with an additional pilot of 9 reasoning models. Our key findings are:

1. **Fine-tuning delivers genuine improvement**: Re-evaluating the 2025 GPT-3.5-FT baseline on our canonical test set confirms 88.0% SYL_R%, while GPT-4.1-mini-FT achieves 93.9%, a +5.9pp improvement at a fraction of the cost of GPT-4.1-FT (93.3%). Performance saturates at the mini tier, while GPT-4.1-nano-FT (34.8% SYL_EM%) reveals a minimum capacity threshold.
2. **Format-tolerant metrics change the narrative**: BND_F1 scores reveal that zero-shot models possess substantial segmentation competence (up to 86.8%) masked by low SYL_EM% scores, suggesting that LLMs have implicit phonological knowledge that fine-tuning helps format correctly.
3. **Reasoning models struggle**: In a pilot study (N=10), all 9 thinking/reasoning models achieve 0% SYL_EM%, failing through three distinct mechanisms: output truncation (reasoning consumes the token budget), empty or leaked responses (models deliberate extensively without converging), and incorrect notation conventions. This suggests that extended deliberation is counterproductive for format-sensitive tasks like scansion.
4. **Portuguese specialization does not transfer to scansion**: general-purpose models outperform Sabiá-family models on this task, but the gap reflects weaker instruction-following (as reported for Sabiá-3 on IFeval (Abonizio et al. 2025)) more than weaker phonology, since their format-tolerant BND_F1 is comparable.
5. **The accessibility gap persists**: The most accurate scansion requires commercial fine-tuned models, while the cheapest alternatives sacrifice substantial accuracy. Fine-tuning open-weight models represents a promising but underexplored pathway.

Future work should explore fine-tuning open-weight models (e.g., LLaMA, Gemma) on the Aoidos corpus to evaluate whether the fine-tuning accuracy observed with proprietary models transfers to open architectures. A complementary direction is token-free, character-level modeling (Belouadi and Eger 2023; Yu et al. 2024): because the dominant errors in our results lie in syllable boundaries and metaplasms, exactly where subword tokens misalign with poetic syllables, such architectures could improve segmentation beyond what conventional fine-tuning achieves. With a reliable scansion module in place, the natural next step is meter-constrained verse generation, where a scansion model serves as a metrical filter or reward signal for a generator, enabling the production of formally valid poetry.

Beyond text generation, integrating low-latency speech synthesis with appropriate prosody would enable spoken-verse systems capable of participating in oral poetic traditions, such as the improvised verse duels of Northeastern Brazil, where meter is not a stylistic choice but a structural requirement.

Appendix: Run-to-Run Consistency

To assess how stable our single-run measurements are at temperature 0.1, we re-ran three representative models, one fine-tuned, one proprietary, and one open-weight, five times each on a fixed 30-verse subsample, and measured run-to-run agreement (Table 4). The fine-tuned model is nearly deterministic: 90% of verses yield byte-identical scansions across all five runs, syllable counts agree on 96.7% and stress patterns on 100%, with only 1.1 distinct outputs per verse on average. Determinism is not uniform across providers, however. Gemma 3 12B remains fairly stable (76.7% identical, 1.3 distinct outputs), whereas Gemini 2.5 Flash varies substantially (33.3% identical, 2.8 distinct outputs per verse) despite the low temperature. Two factors contribute: its few-shot outputs are frequently incomplete, with the truncation point shifting between runs, and the segmentation itself is unstable, since even on the common prefix preceding truncation only 56.7% of verses agree on syllable text. The headline metrics for the fine-tuned models, which are our primary results, are therefore highly reproducible, while single-run scores for high-variance proprietary models should be read as point estimates carrying run-to-run uncertainty.

Table 4: Run-to-run consistency at temperature 0.1: each model scanned the same 30 verses five times. *Exact* = identical output across all five runs; *Count* and *Stress* = identical syllable count and stress pattern across runs; *Distinct* = mean number of distinct outputs per verse (1 = fully deterministic).

Model	Exact%	Count%	Stress%	Distinct
GPT-4.1-mini-FT	90.0	96.7	100.0	1.1
Gemini 2.5 Flash (FS)	33.3	56.7	33.3	2.8
Gemma 3 12B (FS)	76.7	93.3	86.7	1.3

References

Abonizio, H.; Almeida, T. S.; Laitz, T.; Junior, R. M.; Bonás, G. K.; Nogueira, R.; and Pires, R. 2025. Sabiá-3 technical report. arXiv preprint arXiv:2410.12049.

Agirrezabal, M.; naki Alegria, I.; Arrieta, B.; and Hulden, M. 2012. Finite-state technology in a verse-making tool. In *Proceedings of the Finite State Methods and Natural Language Processing Conference*, 35–39.

Agirrezabal, M.; Astigarraga, A.; Arrieta, B.; and Hulden, M. 2016. ZeuScansion: A tool for scansion of English poetry. *Journal of Language Modelling* 4(1):3–28.

Agirrezabal, M. 2017. *Automatic Scansion of Poetry*. Ph.D. Dissertation, University of the Basque Country (UPV/EHU).

Ali, M. S. 1999. *Versificação Portuguesa*. São Paulo: EDUSP.

Almeida, T. S.; Abonizio, H.; Nogueira, R.; and Pires, R. 2024. Sabiá-2: A new generation of Portuguese large language models. arXiv preprint arXiv:2403.09887.

Anthropic. 2026. The Claude model card and system prompt. Accessed: February 2026. `https://docs.anthropic.com/en/docs/about-claude/models`.

Belouadi, J., and Eger, S. 2023. ByGPT5: End-to-end style-conditioned poetry generation with token-free language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7364–7381.

Bisol, L. 2005. *Introdução a estudos de fonologia do português brasileiro*. Porto Alegre: EDIPUCRS, 4 edition.

Bobenhausen, K. 2011. The Metricalizer: Automated metrical markup of German poetry. In Küper, C., ed., *Current Trends in Metrical Analysis*. Frankfurt am Main: Peter Lang. 119–131.

Brown, T., et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems* 33:1877–1901.

Caldas, P. 2021. Ano passado eu morri, mas esse ano eu não morro: verso absurdo une paraibanos a novas gerações. `https://g1.globo.com/pb/paraiba/noticia/2021/01/31/ano-passado-eu-morri-mas-esse-ano-eu-nao-morro-verso-absurdo-une-paraibanos-a-novas-geracoes.ghtml`. Accessed: 2024-02-19.

Chakrabarty, T.; Padmakumar, V.; and He, H. 2022. Help me write a poem: Instruction tuning as a vehicle for collaborative poetry writing. arXiv preprint arXiv:2210.13669.

Chatzikyriakidis, S., and Natsina, A. 2026. LLMs got rhythm? Hybrid phonological filtering for Greek poetry rhyme detection and generation.

Colton, S.; Goodwin, J.; and Veale, T. 2012. Full-FACE poetry generation. In *Proceedings of the 3rd International Conference on Computational Creativity (ICCC)*, 95–102.

Corrêa, N. K.; Falk, S.; Fatimah, S.; Sen, A.; and Oliveira, N. D. 2024. TeenyTinyLlama: open-source tiny language models trained in Brazilian Portuguese. *Machine Learning with Applications* 16:100558.

Corrêa, N. K.; Sen, A.; Falk, S.; and Fatimah, S. 2025. Tucano: Advancing neural text generation for Portuguese. *Patterns* 6(11):101325.

de Araújo, P. A. M., and Mamede, N. J. 2002. Classificador de poemas. In *Conferência Científica e Tecnológica em Engenharia*.

de la Rosa, J.; Álvaro Pérez; Hernández, L.; Ros, S.; and González-Blanco, E. 2020. Rantanplan, fast and accurate syllabification and scansion of Spanish poetry. *Procesamiento del Lenguaje Natural* 65:83–90.

de la Rosa, J.; Álvaro Pérez; de Sisto, M.; Hernández Lorenzo, L.; Díaz, A.; Ros, S.; and González-Blanco, E. 2023. Transformers analyzing poetry: Multilingual metrical pattern prediction with transformer-based language models. *Neural Computing and Applications* 35:18171–18176.

DeepSeek-AI. 2024. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*.

Frolov, D. 2000. *Classical Arabic Verse: History and Theory of ʿArūḍ*, volume 21 of *Studies in Arabic Literature*. Leiden: Brill.

Gervás, P. 2000. A logic programming application for the analysis of Spanish verse. *Computational Linguistics and Intelligent Text Processing* 330–344.

Google DeepMind. 2024. Gemma: Open models based on Gemini research and technology. *arXiv preprint arXiv:2403.08295*.

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)*, 524–533.

Hartman, C. O. 2005. The scandroid 1.1. Software for automated metrical scansion of English verse, `https://academic.hartman.digital.conncoll.edu/Programs.htm`.

Hayward, M. 1996. Analysis of a corpus of poetry by a connectionist model of poetic meter. *Poetics* 24(1):1–11.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Koziev, I. 2025. Automated evaluation of meter and rhyme in Russian generative and human-authored poetry. arXiv preprint arXiv:2502.20931.

Lau, J. H.; Cohn, T.; Baldwin, T.; Brooke, J.; and Hammond, A. 2018. Deep-speare: A joint neural model of poetic language, meter and rhyme. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 1948–1958.

Logan, H. M. 1988. Computer analysis of sound and meter in poetry. *College Literature* 19–24.

Marco Remón, G., and Gonzalo, J. 2020. Automatic scansion of Spanish poetry without syllabification. *arXiv preprint arXiv:2012.12799*.

Meta AI. 2025. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. `https://ai.meta.com/blog/llama-4-multimodal-intelligence/`.

Mistral AI. 2025. Introducing Mistral 3. `https://mistral.ai/news/mistral-3`.

Mittmann, A., and Maia, S. R. 2017. Análise comparativa entre escansões manual e automática dos versos de Gregório de Matos. *Texto Digital* 13(1):157–176.

Mittmann, A. 2016. *Escansão automática de versos em português*. Ph.D. Dissertation, Universidade Federal de Santa Catarina, Centro Tecnológico.

Navarro-Colorado, B.; Ribes Lafoz, M.; and Sánchez, N. 2016. Metrical annotation of a large corpus of Spanish sonnets: Representation, scansion and evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, 4360–4364.

Navarro-Colorado, B. 2017. A metrical scansion system for fixed-metre Spanish poetry. *Digital Scholarship in the Humanities* 33(1):112–127.

Oliveira, H. G. 2012. PoeTryMe: a versatile platform for poetry generation. In *Proceedings of the ECAI 2012 Workshop on Computational Creativity, Concept Invention, and General Intelligence (C3GI)*.

Oliveira, H. G. 2017. A survey on intelligent poetry generation: Languages, features, techniques, reutilisation and evaluation. In *Proceedings of the 10th International Natural Language Generation Conference*, 11–20.

Oliveira, H. G. 2024. Automatic generation of creative text in Portuguese: an overview. *Language Resources and Evaluation* 58(1):7–41.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ormazabal, A.; Artetxe, M.; Agirrezabal, M.; Soroa, A.; and Agirre, E. 2022. PoeLM: A meter- and rhyme-controllable language model for unsupervised poetry generation. arXiv preprint arXiv:2205.12206.

Pires, R.; Abonizio, H.; Almeida, T. S.; and Nogueira, R. 2023. Sabiá: Portuguese large language models. In *Intelligent Systems: 12th Brazilian Conference, BRACIS 2023, Proceedings, Part III*, volume 14197 of *Lecture Notes in Computer Science*, 226–240. Springer.

Valença, A., and Calegario, F. 2025. Experimenting with large language models for poetic scansion in Portuguese: A case study on metric and rhythmic structuring. In *Proceedings of the 16th International Conference on Computational Creativity (ICCC)*.

Walsh, M.; Preus, A.; and Antoniak, M. 2024. Sonnet or not, bot? Poetry evaluation for large models and datasets. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 15568–15603.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2022. Chain-of-thought prompting elicits reasoning in large language models. arXiv preprint arXiv:2201.11903.

Yu, C.; Zang, L.; Wang, J.; Zhuang, C.; and Gu, J. 2024. 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)*, 315–325.