Prompting the Muse: Generating Prosodically-Correct Latin Speech with Large Language Models

Michele Ciletti
University of Foggia / Via Arpi 176, 71121 Foggia (FG), Italy
michele_ciletti.587188@unifg.it

Abstract

This paper presents a workflow that compels an audio-enabled large language model to recite Latin poetry with metrically accurate stress. One hundred hexameters from the *Aeneid* and the opening elegiac *epistula* of Ovid's *Heroides* constitute the test bed, drawn from the Pedecerto XML corpus, where ictic syllables are marked. A preprocessing pipeline syllabifies each line, converts alien graphemes into approximate English–Italian counterparts, merges obligatory elisions, adds commas on caesurae, upper-cases every ictic syllable, and places a grave accent on its vowel. Verses are then supplied, one at a time, to an LLM-based Text-to-Speech model under a compact system prompt that instructs slow, articulated delivery. From ten stochastic realisations per verse, a team of Latin experts retained the best; at least one fully correct file was found for 91% of the 200 lines. Upper-casing plus accent marking proved the strongest cue, while hyphenating syllables offered no benefit. Remaining errors cluster around cognates where the model inherits a Romance or English stress template. The corpus of validated audio is openly released on Zenodo, opening avenues for pedagogy, accessibility, and prosodic research.

1 Introduction

Latin prosody, at its core, is the systematic study of Latin poetry, particularly its laws of meter. Unlike English poetry, which relies on the alternation of stressed and unstressed syllables to create rhythm, classical Latin meter operates on a quantitative rhythm, determined by the arrangement of long and short syllables. The very term "prosody" finds its origins in the Greek word *prosoidia*, which initially signified a song sung to music or the specific pronunciation of a syllable.

Whereas handbooks faithfully describe reconstructed prosodical pronunciations, convincing spoken renditions accessible to learners remain scarce. Neural text-to-speech has closed the quality gap for modern languages, yet Latin remains marginal: the models lack training data and frequently transplant English or Romance stress patterns.

Recent work in prosody editing offers an alternative. FastSpeech-type architectures expose duration, pitch, and energy predictors that can be edited after inference (Ren et al., 2020; Lam et al., 2025). Large language models with direct audio decoders add the possibility of steering pronunciation through plain text prompts, avoiding re-training. Their potential for historical languages has scarcely been explored.

The present study therefore asks whether prompt engineering, reinforced by symbolic prosodic annotation, is enough to make a general-purpose LLM read Latin verse with metrically correct stress.

2 Theoretical Background

2.1 Latin Prosody

Classical verse rests on the opposition of long and short syllables, organised into metrical feet and regulated by fixed caesural patterns (Fortson IV, 2011). Quantity derives from vowel length and from consonantal environment, yet several phenomena blur the rule set: *muta cum liquida* allows optional resolution, while pervasive elision removes entire syllables at morpheme borders. Quantitative rhythm therefore resists categorical annotation; even the primary grammarians disagree in boundary cases. Because no contemporary acoustic evidence survives, phonological reconstruction must triangulate between Roman orthography, comparative Romance data, metrical practice, and prescriptive grammars (Allen, 1989). In practice, full reconstruction of absolute vowel length remains tentative. Modern pedagogy often replaces quantity with stress-based recitation, although stress in Latin is governed by its own moraic calculus. Any synthetic-speech system must decide which of these competing principles to privilege.

2.2 Digital Latin: Corpora, Annotation, and Prosodic Tooling

Over three decades, Latin has moved from an almost text-only digital presence to a language with a modest but growing NLP stack (Riemenschneider and Frank, 2023). Tokenisers, lemmatisers, and treebanks are available through resources such as CLTK (Johnson et al., 2021), Stanza (Qi et al., 2020), and the Universal Dependencies Latin collections (De Marneffe et al., 2021). Prosodic annotation, however, remains rarer. Pedecerto (Colombi et al., 2011) annotates circa 244,000 dactylic lines from Musisque Deoque (Mastandrea et al., 2007), returning syllabification, quantity, foot structure, and caesurae. Its XML export supplied the gold data used in the present study. Other scanners address particular metres: the CLTK modules for hexameter and hendecasyllable (Johnson et al., 2021), Anceps for trimeters (Fedchin et al., 2022), and Loquax for quantitative syllabification and IPA transliteration (Court, 2025).

2.3 Large Language Models and Prompt-Based Prosody

Large language models trained on audio-text pairs have begun to encode prosodic regularities that can be elicited by prompt design. VALL-E (Chen et al., 2024) and ZM-Text-TTS (Saeki et al., 2023) exploit massive multilingual corpora; their output retains speaker identity and sentence melody yet shows limited control over metre (Lam et al., 2025). The innovation proposed here inverts the usual pipeline: instead of sampling latent style tokens, we preprocess the poetic text, marking ictic positions and supplying approximate phonology in an orthography already mastered by the model (chiefly English with occasional Italian spellings for /u/ and palatals). At synthesis time those stress markers override default duration predictors, favouring long phones in ictic slots and shortened ones elsewhere. This approach follows the philosophy of PRESENT—prosody is steered through the input representation, not through additional parameters—yet applies it to classical verse rather than conversational prose.

2.4 Pedagogical and Inclusive Perspectives

Audio renditions of Latin verse remain an expensive commodity, created by a handful of trained classicists. Automated generation promises open collections usable in language teaching, literary analysis, and accessibility contexts. Recent surveys in Digital Humanities stress the need for sharable, standardised, and FAIR corpora of recitations (De Sisto et al., 2024). By leveraging TTS engines and releasing the aligned text–audio pairs, the project aims to partially answer that call. Moreover, directing attention to stress rather than absolute quantity lowers the entry barrier for learners whose first language lacks phonemic length, while retaining a recognisable metrical pulse, in accordance with teaching standards across the world.

3 Methodology

3.1 Corpora and Metrical Annotation

The experiments draw on two well-known Latin texts: the opening one hundred hexameter lines of Vergil's *Aeneid* and the first elegiac *epistula* of Ovid's *Heroides*. Together they furnish examples of the two metres most frequently met in both school curricula and introductory prosody courses. A dactylic hexameter line consists of six feet, each prototypically realised as a long–short–short (dactyl, D) or long–long (spondee, S) sequence; the fifth foot is normally a dactyl and the sixth is a spondee whose final syllable is anceps. The elegiac couplet pairs such a hexameter with a dactylic pentameter, divided by a diaeresis after the third arsis; in practice the pentameter is felt as two hemiepes with obligatory caesura.

Machine-readable scansion came from the Pedecerto project (Colombi et al., 2011). Pedecerto encodes each word with a `sy` attribute that enumerates syllables and marks ictic positions with an upper-case A. A fragment of the XML illustrates the structure:

```
<line name="1" meter="H" pattern="DDSS">
  <word sy="1A1b" wb="CF">Arma</word>
  <word sy="1c2A2b" wb="CF">uirumque</word>
  ...
</line>
```

During import the parser retained verse boundaries, foot patterns, ictus markers, word-boundary flags, and elision hints.

3.2 Text Preparation Pipeline

Each line was passed through an iterative preprocessing routine and immediately spoken by a synthesis model; Latinists then annotated pronunciation errors, after which the routine was adjusted. Syllabification relied on the Classical Language

Toolkit, whose rule-based engine already covers enclitics and diphthongs (Johnson et al., 2021). A grave accent was placed over the vowel of every ictic syllable and the entire syllable was upper-cased. Words forming obligatory elision were merged (`quoque et → quoquet`) in accordance with the Pedecerto `wb` attribute. Caesurae were rendered by a comma, but only when the manuscript transmitted no other punctuation at that position; this decision proved particularly useful for pentameter lines, where the pause after the third arsis is nearly fixed. Trials in which syllables were separated by hyphens (`ar-ma vi-rum-que`) showed no measurable benefit and were dropped.

Orthographic substitution aimed at a rough classical pronunciation that modern English or Italian acoustic models could approach. Stops before front vowels were written `k` instead of `c`; `qu` became `kw`; `ae` and `oe` became `ai` and `oi`; `ge` and `gi` were expanded to `ghe` and `ghi`.

Because long contexts tended to blur prosodic control, each verse was spoken in isolation. A verse forms a minimal rhythmic unit whose internal pattern must remain coherent, whereas inter-verse junctures tolerate short pauses.

3.3 Speech Synthesis Experiments

Two families of systems were compared. Conventional sequence-to-sequence TTS engines, such as `Tacotron 2` (Shen et al., 2018), `Kokoro` (Hexgrad, 2025), `tts-1` (OpenAI, 2025b), and `tts-1-hd` (OpenAI, 2025a), could not ingest elaborate instructions; their output mis-stressed Latin loans that resemble high-frequency English forms and showed erratic vowel quantity. Large language models with integrated audio decoders performed better, presumably because the system prompt can impose prosodic policy. Models in the GPT-4o and Gemini lines, namely `gpt-4o-mini-tts` (Hurst et al., 2024), `gemini-2.5-pro-preview-tts` (Gemini Team, Google, 2025), and `gemini-2.5-flash-preview-tts` (Gemini Team, Google, 2025), were tested by generating a subset of ten randomly sampled verses several times. A qualitative analysis deemed that `gpt-4o-mini-tts` delivered the most consistent rhythm and segmental clarity, while also being the only model capable of reliably outputting an accurate version of each test verse.

Experiments with original Latin text as the input failed, with no model capable of consistently generating accurate pronunciations of each test verse.

Prompt engineering proceeded from a verbose style sheet to a compact directive. Lengthy system prompts improved intonational contour but occasionally confused stress placement. The final prompt retained only three imperatives: speak slowly, articulate every syllable, obey the marked stresses. Repeating the fully processed verse inside the prompt, exactly as the model should pronounce it, brought an unexpected improvement, perhaps because the acoustic decoder aligns its plan with the visible text.

As LLMs incorporate stochastic sampling, pronunciation varies across runs. For each verse ten realisations were generated. When specialists reviewed the set, at least one rendition met the acceptance threshold in 91 percent of lines. Most remaining errors involved lexical interference from Romance or English cognates; for instance, the word `passus` from the Aeneid's fifth line emerged as `pàssus` rather than the required `passùs`. Re-spelling the stressed vowel (`passùus`) in the prompt usually resolved the problem, though this fix was applied sparingly, since excessive vowel doubling sometimes misled the model elsewhere in the line.

Sequences with dense stress, such as spondaic clusters, challenged the model, as did runs of elided vowels or complex consonant groups. These limitations are examined in Section 5.

3.4 Expert Evaluation Protocol

Three scholars of Latin phonology, none involved in system development, evaluated every candidate recording. Errors were marked on a verse basis and classified as segmental, stress, elision, or pacing. Feedback was returned after each experimental cycle, leading to successive refinements of preprocessing and prompts until the acceptance rate stabilised.

3.5 Dissemination of Audio Material

The highest-ranked file for each verse was retained. Verses were concatenated with 800 ms silences, yielding two continuous recitations that mirror performance practice yet preserve per-line rhythmic autonomy. Waveform-level normalisation ensures homogeneous loudness. The corpus has been deposited on Zenodo (Ciletti, 2025) under a Creative Commons Attribution 4.0 license. (Commons, 2016)

Metre	Lines	Lines with at least one correct realisation
Hexameter	158	91.1%
Pentameter	58	91.4%
Total	216	91.2%

Table 1: Overview of the obtained Latin verse recordings.

4 Results

4.1 Quantitative Assessment

The evaluation covered 216 autonomous lines, of which 158 hexameters and 58 pentameters. Ten recordings were generated for every line, yielding two thousand candidate files. Table 1 reports acceptance rates after expert screening. The final system prompt is as follows:

> This is a Latin poetical verse. Pronounce it rhythmically, slowly and with emphasis, articulating each syllable and correctly stressing them. Pronounce it like this: [pre-processed verse]

Incorrect verses fell into four categories: segmental substitutions, misplaced ictus, elision failure, and pacing anomalies. Inter-annotator agreement on the five-way label reached $\kappa = 0.79$ for hexameter and $\kappa = 0.84$ for pentameter. Most of the disagreements arose from cases where two different types of errors overlapped (such as incorrect stress paired with mispronounced words). After several rounds of discussion, the annotators agreed on the most prominent error for each verse, and all discrepancies were resolved.

The overall accuracy of the model stood at 59.03%.

4.2 Effect of Preprocessing Variants

Ablation tests, run on a ten-line subset to contain annotation effort, show that three operations account for most of the gain over a plain graphemic baseline:

- Upper-casing and accenting the ictic syllable considerably reduced stress errors;
- Orthographic substitution of `c`/`qu`/`ae`/`oe` and palatal stops diminished segmental errors;
- Explicit commas on caesura lowered pacing mistakes, especially in pentameters.

Conversely, syllable hyphenation had negligible impact, while long system prompts improved intonation without improving segmental or stress accuracy. These findings corroborate earlier observations by Lam et al. (2025) that explicit duration–pitch instructions dominate hidden stylistic embeddings in LLM-based TTS.

4.3 Listening Quality

Mean opinion scores were collected from fourteen external listeners familiar with Latin recitation but naïve to the study. They judged naturalness and metrical fidelity on a five-point scale. Best-of-ten selection reached 4.1 ± 0.6 for hexameter and 3.9 ± 0.7 for pentameter. Ratings drop by roughly one point when a random sample rather than the best file is played, reflecting the intrinsic variance of stochastic decoding.

5 Conclusions and Outlook

The workflow demonstrates that a contemporary audio-enabled large language model, guided by minimal yet well-targeted textual cues, can read classical Latin verse with a promising degree of prosodic correctness. Stress salience carried by case-shift and diacritic proved a stronger cue than any attempt at modelling moraic weight directly, an outcome consistent with linguistic evidence on the rhythmical importance of stress in Latin poetry (Pawlowski and Eder, 2001). Segmental confusion arises chiefly from orthographic overlap with Italian and English; paradoxically, rare or morphologically opaque words are rendered more faithfully because no competing template exists in the model's training distribution.

5.1 Future Work

Two lines of research appear promising. First, coupling the current prompt-based strategy with the controllable duration and energy interfaces available in FastSpeech-type decoders (Ren et al., 2021) may supply the missing quantitative layer. Second, training a lightweight alignment model on our validated recordings would allow deterministic selection rather than trial-and-error sampling. Beyond technology, the public release on Zenodo of both source XML and mastered audio will facilitate studies in metrics, second-language acquisition, and accessibility. The same pipeline applies, *mutatis mutandis*, to other Greco-Roman metres, to post-classical accentual hymns, and even to vernacular verse traditions where scholarly recordings are scarce. Furthermore, a dataset of manually curated audio files could be promising for the purpose of fine-tuning smaller, open-source text-to-speech models.

Limitations

The system remains probabilistic. A user must be willing to request several readings and to curate the output manually. Dense spondaic passages, intricate elisions, and clusters such as ctn or gns still trigger mis-timed syllable nuclei. Quantity is approximated through pacing alone; true heavy-light contrast, audible as durational ratio, is not yet guaranteed. Finally, the present study uses a single North-Atlantic vocal profile, whereas pedagogy would profit from multiple voices and speaking rates. Accurate results remain dependent on manual verification and prompt adjustments for specific verses; improvements are necessary to fully automate the pipeline and enhance its productivity.

Acknowledgments

The author thanks the entire Pedecerto team for annotating and sharing their XML scansions. Sincere gratitude is extended to the CLTK community, whose open-source tools simplified syllabification and phonological checks. Colleagues at the University of Foggia donated hours to the auditory review of candidate recordings; their expertise shaped both the preprocessing rules and the acceptance thresholds. Any remaining inaccuracies are the sole responsibility of the author.

References

W Sidney Allen. 1989. *Vox Latina: a guide to the pronunciation of classical Latin*. Cambridge University Press.

Sanyuan Chen, Shujie Liu, Long Zhou, Yanqing Liu, Xu Tan, Jinyu Li, Sheng Zhao, Yao Qian, and Furu Wei. 2024. Vall-e 2: Neural codec language models are human parity zero-shot text to speech synthesizers. *arXiv preprint arXiv:2406.05370*.

Michele Ciletti. 2025. Veras audire et reddere voces: A corpus of prosodically-correct latin poetic audio from large-language-model tts.

Emanuela Colombi, Luca Mondin, Luigi Tessarolo, Andrea Bacianini, Dylan Bovet, and Alessia Prontera. 2011. Pedecerto. *Pedecerto. Metrica Latina Digitale*.

Creative Commons. 2016. Creative commons attribution 4.0 international public license. Accessed: 2025-06-29.

Matthieu Court. 2025. Loquax: Nlp framework for phonology. `https://github.com/mattlianje/loquax`. GitHub repository.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

Mirella De Sisto, Laura Hernández-Lorenzo, Javier De la Rosa, Salvador Ros, and Elena González-Blanco. 2024. Understanding poetry using natural language processing tools: a survey. *Digital Scholarship in the Humanities*, 39(2):500–521.

Aleksandr Fedchin, Patrick J Burns, Pramit Chaudhuri, and Joseph P Dexter. 2022. Senecan trimeter and humanist tragedy. *American Journal of Philology*, 143(3):475–503.

Benjamin W Fortson IV. 2011. Latin prosody and metrics. *A companion to the Latin language*, pages 92–104.

Gemini Team, Google. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. Google DeepMind Technical Report. Version from 2025-06-17.

Hexgrad. 2025. Kokoro-82m (revision d8b4fc7).

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Kyle P Johnson, Patrick J Burns, John Stewart, Todd Cook, Clément Besnier, and William JB Mattingly. 2021. The classical language toolkit: An nlp framework for pre-modern languages. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing: System demonstrations*, pages 20–29.

Perry Lam, Huayun Zhang, Nancy F Chen, Berrak Sisman, and Dorien Herremans. 2025. Present: Zero-shot text-to-prosody control. *IEEE Signal Processing Letters*.

Paolo Mastandrea and 1 others. 2007. Musisque deoque. un archivio digitale di poesia latina, dalle origini al rinascimento italiano.

OpenAI. 2025a. Openai tts-1-hd model documentation. Accessed: 2025-06-29.

OpenAI. 2025b. Openai tts-1 model documentation. Accessed: 2025-06-29.

Adam Pawlowski and Maciej Eder. 2001. Quantity or stress? sequential analysis of latin prosody. *Journal of Quantitative Linguistics*, 8(1):81–97.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*.

Frederick Riemenschneider and Anette Frank. 2023. Exploring large language models for classical philology. *arXiv preprint arXiv:2305.13698*.

Takaaki Saeki, Soumi Maiti, Xinjian Li, Shinji Watanabe, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2023. Learning to speak from text: Zero-shot multilingual text-to-speech with unsupervised text pretraining. *arXiv preprint arXiv:2301.12596*.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. *Preprint*, arXiv:1712.05884.