Umanistica Digitale – ISSN: 2532-8816 – n. 23, 2026: Special Issue
M. Ciletti – *Generating Prosodically Accurate Audio of Latin Poetry with Text-to-Speech Large Language Models: A Computational Workflow*
DOI: http://doi.org/10.60923/issn.2532-8816/23570

Prompting the Muse

Generating Prosodically Accurate Audio of Latin Poetry with Text-to-Speech Large Language Models: A Computational Workflow

Michele Ciletti

Università di Foggia
michele.ciletti@unifg.it

Abstract

While the field of Digital Humanities has successfully established robust infrastructures for the textual analysis of Latin, the auditory dimension of the language is still largely undeveloped. Specifically, the complex quantitative rhythm and intonation of classical poetry cannot be accurately replicated by Text-to-Speech models. This paper presents a computational workflow designed to bridge this gap, leveraging verified metrical data to produce high-fidelity and prosodically accurate audio recordings. By using the structured XML scansions of the *Pedecerto* project, the proposed pipeline employs a rule-based pre-processing routine to convert standard orthography into a phonetic script optimized for acoustic modelling. These adapted texts are then fed into a multimodal Large Language Model, which is steered via in-context prompt engineering to observe syllable quantity, ictus placement, pauses, and possible elision. The technical architecture of this system is detailed, analyzing the specific orthographic interventions and prompts required to overcome the stress-timed bias of contemporary AI models. Finally, the implications of this tool for the wider Digital Humanities ecosystem are discussed, with particular attention to its potential to democratize access to Latin learning, support accessibility, and add new audio layers to existing digital projects and infrastructures.

Keywords: Digital Humanities; Latin Poetry; AI; Computational tools

Sebbene il campo delle Digital Humanities *abbia sviluppato con successo solide infrastrutture per l'analisi testuale del latino, la dimensione sonora della lingua rimane ancora in gran parte inesplorata. In particolare, il complesso ritmo quantitativo e l'intonazione della poesia classica non possono essere riprodotti con accuratezza dai modelli di sintesi vocale (Text-to-Speech). Questo contributo presenta un flusso di lavoro computazionale progettato per colmare tale lacuna, sfruttando dati metrici verificati per produrre registrazioni audio ad alta fedeltà e prosodicamente accurate. Utilizzando le scansioni metriche strutturate in XML del progetto* Pedecerto*, la pipeline proposta impiega una procedura di pre-elaborazione basata su regole per convertire l'ortografia standard in una trascrizione*

Copyright © 2026 The Author(s)
The text in this work is licensed under the Creative Commons BY License.
https://creativecommons.org/licenses/by/4.0

fonetica ottimizzata per la modellazione acustica. Questi testi adattati vengono quindi forniti a un Large Language Model multimodale, guidato tramite prompt engineering in-context *a rispettare la quantità sillabica, la posizione dell'ictus, le pause e i fenomeni di elisione. L'architettura tecnica del sistema viene descritta in dettaglio, analizzando gli interventi ortografici specifici e le strategie di* prompting *necessarie per superare il bias verso il ritmo accentuativo tipico dei modelli di intelligenza artificiale contemporanei. Infine, vengono discusse le implicazioni di questo strumento per l'ecosistema più ampio delle* Digital Humanities*, con particolare attenzione al suo potenziale nel democratizzare l'accesso allo studio del latino, supportare l'accessibilità e aggiungere nuove dimensioni sonore ai progetti e alle infrastrutture digitali esistenti.*

Parole chiave: Digital Humanities; Poesia latina; IA; Strumenti computazionali

1. Introduction

Over the last years, the digitization of Classical heritage and the development of computational tools to analyze it has achieved a level of maturity and granularity that was once difficult to imagine. Through collective efforts, we now possess extensive, searchable repositories of text, morphological analyzers, interconnected knowledge graphs that link authors, works, and linguistic phenomena across centuries [13]. Yet, amidst this abundance of textual and visual data, the auditory dimension is still underrepresented.

The Latin language cannot be thoroughly understood without considering its oral and aural phenomena. This is particularly true for its poetry, which was composed to be performed and heard, apart from being read. However, the current digital infrastructure for Latin is almost entirely devoid of audio: a user can query a database to find every instance of a dactylic hexameter in Vergil or retrieve the morphological analysis of a specific lemma, but the auditory realization of that data is usually left to the user's imagination.

This situation is complicated by the fact that the unique prosodical nature of Latin poetry, which operates on principles that are fundamentally different from those of most modern Western languages, makes it hard to faithfully capture and reproduce. Classical Latin meter is quantitative, relying on the duration of syllables, and more specifically on the opposition between heavy (long) and light (short) elements. Let's take the hexameter, the most common verse structure in Latin epic, as an example: it is, in its simplest form, made up of five repeating patterns of long-short-short syllables (dactyls), followed by a final long-long (spondee) or long-short (trochee) pattern. However, most of these dactyls may be substituted with spondees at the discretion of the author, creating a relative ambiguity. From a rhythmical perspective, the misalignment between the metrical ictus and the natural word accent creates a further difficulty [6]. Furthermore, phenomena such as elision, where vowels at word boundaries merge or disappear, and the placement of conventional pauses (*caesurae*), which dictates the phrasing and breath of the line, are strictly codified.

For modern students or researchers, mastering this system can prove to be a challenge. The standard method of teaching scansion involves marking texts with macrons and breves, which is a certainly useful but purely visual abstraction. Without a reliable acoustic model, self-taught learners can struggle with grasping the specific rhythm of a verse, and this barrier is even more exclusionary for visually impaired scholars, for whom the "visual-only" approach to digital

metrics makes the music of the verse inaccessible. While high-quality recordings curated by experts exist, they are rare, time-consuming to produce, and cover only a minuscule fraction of the vast collection of poetry we possess. This is the rationale behind the proposal of a scalable, consistent solution to try to "hear" the data we have digitized.

Until very recently, bridging the gap between text and prosodically accurate audio was a technological impossibility for low-resource languages like Latin. Traditional Text-to-Speech (TTS) systems, such as those used in commercial assistants, are typically trained on vast datasets of modern English or Romance languages. When forced to read Latin, these models inevitably apply the prosodic rules of their training data, ignoring vowel quantity and misplacing stress [3]. Correcting these errors in earlier architectures required training a new voice model from scratch, a process demanding hours of studio-quality recordings and deep technical expertise that is often beyond the reach of a typical Humanities research team.

The emergence of multimodal Large Language Models (LLMs) has fundamentally changed this landscape. Unlike their predecessors, contemporary models such as OpenAI's GPT-4o [8] or Google's Gemini 2.5 [5] are increasingly capable of understanding and generating audio as a native modality. Because of this shift, in-context steerability is now possible.

Because these models are trained on massive, multilingual corpora, they possess a latent understanding of many different phonological structures. More importantly, they can be directed via natural language instructions, known as system prompts, to modify their delivery in real-time and let them learn new information from the input itself, a process which is called in-context learning [2]. Instead of retraining the entire model, a user can now describe the desired output: instructing the model to speak slowly, to articulate every syllable, or to place emphasis on specific marked vowels. This technique, known as prompt engineering [11], gives a relatively high degree of control even to non-specialized people.

However, LLMs remain probabilistic engines, so their processes are inherently non-deterministic. They are prone to hallucinations [14] in pronunciation, occasionally reverting to English phonemes or ignoring metrical instructions. Therefore, relying solely on a raw model is insufficient for academic purposes. To fully harness this technology, a structured workflow is needed.

The present contribution describes such a workflow. It details a pipeline designed to integrate with the existing ecosystem of Digital Latin tools, specifically leveraging the XML scansions of the *Pedecerto* project[1]. By combining their precise philological data with targeted prompt engineering, this framework compels a general-purpose language model to function as a specialized Latin TTS engine. The following sections will illustrate the architecture of this system, together with its implications and possibilities.

1 https://www.pedecerto.eu/.

2. Pipeline Architecture

Any attempt to automate the recitation of Latin poetry must begin with a recognition of the limits of raw text. A Large Language Model, no matter how sophisticated, cannot deduce the correct metrical quantity of a line solely from a plain string of characters, as the variable nature of syllable weight and metrical ictus often strips away the necessary prosodic information. Therefore, the architecture proposed here, rather than a standalone generative system, is a downstream application that rests upon the foundation of structured, annotated data.

To validate the workflow, a corpus that represents some of the most common verse structures of the Latin canon was selected, spanning the Golden Age of Roman literature. The primary test bed consists of the first one hundred hexameters of Vergil's *Aeneid* and the opening elegiac epistle of Ovid's *Heroides* (116 lines). This selection aims to make the system approach both the relentless, rolling momentum of the epic hexameter and the alternating, asymmetrical pulse of the elegiac couplet, which consists of a hexameter followed by a pentameter. These texts offer a variety of phonotactic challenges, such as heavy spondaic lines and complex elisions that test the boundaries of articulation.

The ground truth for these lines was ingested directly from the *Pedecerto* project. As one of the premier digital archives for Latin poetry, *Pedecerto* provides highly granular XML files where every syllable is indexed, its quantity is defined, and the metrical ictus is explicitly marked. By parsing the *Pedecerto* XML export, the proposed workflow extracts a series of rhythmic features for every verse. By leveraging the Classical Language Toolkit (CLTK) library [9], the system employs a syllabifier and a macronizer, in addition to the obtained metrical information, to identify the syllables a verse is composed of, their quantity, which syllables carry the ictus, where the foot boundaries lie, and which words must be elided to fit the meter. This ensures that the audio generation is based on scholarly consensus and evidence.

Once the metrical data is extracted, it must be translated into a format that an audio-enabled LLM can interpret. Since LLMs are predominantly trained on English and modern Romance languages, they possess a strong "phonetic bias". If presented with the standard Latin orthography of *Cicerō*, a model is likely to default to an English soft "c" (/s/) or an Italianate pronunciation (/ʧ/), ignoring the classical hard stop (/k/). Furthermore, without visual cues, the model will almost certainly apply a stress-based accentual pattern typical of modern speech, disregarding the quantitative ictus of the verse.

To counteract this, the pipeline employs an iterative pre-processing routine that functions as a phonetic translator. This module rewrites the Latin text into a specific pseudo-orthography designed to steer the model's acoustic output. First, the script handles the critical issue of stress: based on the *Pedecerto* markers, every syllable bearing the metrical ictus is capitalized and its vowel is marked with a grave accent (e.g., ÀR in *arma*). This combination of visual cues, capitalization indicating volume or importance, and the diacritic indicating stress, has proven to be the most effective signal for overriding the model's default intonation patterns [3].

Secondly, the pre-processor addresses segmental phonology. To prevent the model from drifting into ecclesiastical or anglicized pronunciation, the script systematically replaces ambiguous graphemes with phonetically clear alternatives. For instance, the letter "c" is rendered as "k" before front vowels to ensure a velar stop; the digraph "qu" is converted into "kw"; and some diphthongs are respelled to approximate their original pronunciation ("ae" becomes "ai" and

"oe" becomes "oi"). Finally, the problem of elision (the blending of vowels across word boundaries) is solved by merging the tokens of the affected words in the text string. If the meter, for instance, requires *monstrum horrendum* to be elided, the prompt fed to the model will present the fused form *monstrhorrendum*. This orthographic fix forces the model to treat the sequence as a single phonological unit, effectively bypassing the pause it would naturally insert between words.

The engine that powers this pipeline is the audio-generation capability of multimodal Large Language Models. In the experiments detailed in the associated case studies [3], OpenAI's GPT-4o series [8] was primarily utilized, specifically the gpt-4o-mini-audio-preview endpoints, which offered a favourable balance between inference speed, cost, and control. However, it is important to note that the workflow described here is model-agnostic.

The field of generative audio is evolving at a rapid pace: at the time of writing, new architectures from competitors such as Google (Gemini 2.5 Pro) [5] and open-weight models emerging on platforms like Hugging Face (such as the Orpheus and Sesame families) are being tested for their suitability in this pipeline. The criteria for selection remain constant: the model must support direct text-to-speech and must be steerable via natural language prompts. The framework is designed to be modular, allowing the LLM component to be swapped out as more capable or more accessible models become available to the Digital Humanities community.

3. Prompt Engineering

A true novelty of using Large Language Models for speech synthesis lies in their responsiveness to custom instructions. This capability rests on the principle of in-context learning [2], where a model adapts its behaviour based on the examples and directives provided within the input window itself. For the Digital Humanities, this is a considerable opportunity: traditionally, adapting a speech synthesis engine to a new language or a specific historical accent required fine-tuning the model's weights, which is a computationally expensive process that demands vast datasets and significant GPU resources. By contrast, prompt engineering allows us to partially steer the model's output simply using natural language. This also aligns with the principles of Minimal Computing [7], thus helping reduce the barriers to entry and allowing users to generate high-quality results without specialized hardware or Computer Science expertise.

The prompt serves as the style guide for the model. The development of the system prompt described in this workflow was iterative: several different strategies were tested, from longer and detailed ones to more concise options. The final system prompt, deemed to be the most effective in previous studies [3], passes the following instructions to the model: "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]".

One key discovery in this process was the utility of the instruction "slowly". In early tests, models attempting a natural conversational pace often approximated the pacing of certain sequences or blurred complex consonant clusters. By explicitly asking for a slower delivery, some momentum can be sacrificed in exchange for significantly higher segmental precision. However, this is a variable, not a constant; the prompt is essentially a set of open parameters. A user wishing for a

more rapid, dramatic reading could simply remove the tempo instruction or replace it with an emotive descriptor (e.g., "with urgency"). Thanks to this relative flexibility, the framework can serve different pedagogical or aesthetic needs without altering the underlying code. Furthermore, all models react differently to specific prompting styles: while the aforementioned strategy was proven to be highly effective for GPT-4o, other models, especially belonging to different families, may benefit from entirely new instructions. Because of this, research on efficient prompting approaches continues.

Despite the structured pre-processing and careful prompting, current LLMs remain probabilistic engines. During the experiments, it was found that lowering the temperature (the parameter controlling randomness) to near-zero, which is a standard practice in scientific settings to ensure replicability [1], often caused the model to fixate on erroneous pronunciations, repeating the same mistakes in a loop. Conversely, a moderate temperature setting allowed the model to explore different acoustic realizations. Therefore, the workflow should operate on a multi-generation basis: each verse is synthesized in isolation, typically around ten times, to produce a collection of variants.

This probabilistic nature necessitates a human-in-the-loop [12] approach. The machine does the heavy lifting of speech production, but the scholar remains the ultimate judge of the produced output. In the creation of the *Veras Audire et Reddere Voces* corpus, the first major output of this pipeline, every line was audited by experts [4]. An analysis of this data showed that while the model achieved a high success rate (finding at least one fully correct rendition for over 90% of lines), errors were still common, and they mainly clustered, predictably, around Latin words with close cognates in English or Romance languages. For instance, the model frequently attempted to apply an English stress pattern to the word *passus*, stressing its first syllable (mimicking the English *pass* or the Italian *passo*) even when instructed to place an ictus on the second one. Interestingly, it was found that purely orthographic hacks were often sufficient to correct these biases: doubling the vowel of the stressed syllable (e.g., rewriting *passus* as *passuus* or *passuuus*) proved a powerful solution, effectively giving more weight to the syllable in the model's attention mechanism. This also reflects findings in similar workflows designed for different languages and contexts, such as PRESENT [10]. Doing so, any remaining errors in the experimental corpus were fixed. Conversely, other linguistic strategies, such as hyphenating syllables (*pas-sus*), yielded no measurable benefit.

It is important to acknowledge that prompt engineering is an empirical methodology, highly dependent on the specific model version in use. As newer architectures with different training data distributions are released, the optimal prompt will likely evolve. Thus, the prompt and approaches provided in this framework should be viewed as a baseline, a proven starting point from which the community can experiment and refine.

4. Usage Scenarios and Case Studies

The workflow described above is intended to become a practical instrument designed to serve the diverse needs of the Latin and Digital Humanities communities. Several new avenues for research, teaching, and accessibility can be opened.

The most immediate application of this pipeline is the creation of static and verified audio corpora. As a proof of concept, *Veras Audire et Reddere Voces* was released, a collection of 216 fully audited lines from Vergil and Ovid [4]. This *corpus* simply aims to demonstrate the basic functioning of the workflow: hundreds of lines can be processed and synthesized in a matter of minutes. For digital libraries and infrastructures, this can be a scalable path to implementing acoustic data in their collections. Instead of hosting silent text, archives could provide an accompanying audio track for every work (or some of them), allowing users to toggle between reading the scansion and hearing it. Furthermore, the corpus can serve as an approved test bed for future research projects, as will be proposed in the next section.

Beyond research and archives, this tool was created with educators in mind: in a classroom setting, the ability to generate custom audio on demand can be valuable and efficient. A teacher explaining Latin poetry could generate different examples of specific phenomena, or even produce incorrect versions to test students' listening skills. The flexibility of the prompt also allows for stylistic variation, so teachers can request slower, hyper-articulated versions for beginners, or faster recitations for advanced students. This would allow educators to create bespoke listening comprehension exercises tailored to their specific curriculum.

Another high-potential application of this technology lies in accessibility. For self-taught learners without access to expert tutors, fully grasping the rules of Latin meter can prove to be difficult. An on-demand generator provides a feedback loop that textbooks lack, allowing the student to verify their own recitations against a correct auditory model. Furthermore, for visually impaired persons, for whom the visual scansion system is inherently exclusionary, this workflow could offer a new alternative. Even just for students who simply prefer to learn by listening, such an approach could be an accessible way to understand poetry more comfortably.

5. Integration, Extensibility, and Future Perspectives

The framework presented here is the beginning of a much larger project: even though the current iteration relies on commercial APIs for high-quality generation, namely those provided by OpenAI, the ultimate goal is to move this capability into the open scholarly infrastructure. The dataset of validated audio produced by this pipeline, in the future, will serve as part of the training material for smaller, specialized models. By fine-tuning open-source architectures (such as the previously mentioned Orpheus and Sesame) on this prosodically accurate data, the community could benefit from lightweight and offline solutions that run directly in a browser or a local machine. This would democratize access further, removing most of the actual cost barriers and ensuring long-term preservation independent of corporate APIs.

The logic of the pipeline, which ultimately consists of converting text into a phonetic script and steering a model via natural language prompts, is also inherently transferable: the same principles can be applied to virtually any other language. With appropriate adjustments to the transliteration module, this workflow could be tested on Ancient Greek, Classical Arabic, or Old Occitan, provided that high-quality digital scansions are available to guide the process. Furthermore, the flexibility of the prompt could allow dialectal experiments even within the same language; users, for instance, could theoretically instruct the model to adopt an Ecclesiastical pronunciation for medieval Latin texts.

Metrically speaking, the possibilities are equally broad. While these initial experiments focused on the dactylic hexameter and elegiac couplet, the rigid momentum of Roman comedy's iambic senarius suggests it would be an excellent candidate for this approach, possibly yielding even better results. Finally, it would be useful to integrate these audio generation capabilities with existing projects, such as the *LiLa Knowledge Base*[2] or the *Musisque Deoque* project[3], embedding this generative capacity into shared digital environments. The main goal would be to enhance these repositories with a new acoustic dimension.

6. Conclusion

The digitization of Latin has reached impressive heights, as evidenced by multiple successful projects and infrastructures. One of the next frontiers is expanding this process beyond text, in a multimodal perspective. The workflow described in this paper aims to show that this recovery does not necessarily need new technological breakthroughs; the required tools are already in our hands. By pairing the philological accuracy of established metrical databases with the generative flexibility of Large Language Models, it is now possible to leverage a reliable and reproducible method for creating prosodically faithful audio.

This synergy could bring new, exciting possibilities for future research in historical languages. It seems evident that Large Language Models represent the next frontier in Computational Linguistics as a whole. In the case of Latin prosody, we have seen how a general-purpose model, when properly guided, can transcend its training limitations to reproduce the standardized patterns of quantity and ictus. As this technology matures and integrates with the wider ecosystem of Digital Humanities, it could enrich our interaction with the classical world, enabling new multisensory experiences in the study of Latin.

References

[1] Abdurahman, Suhaib, Alireza Salkhordeh Ziabari, Alexander K. Moore, Daniel M. Bartels, and Morteza Dehghani. 2025. "A primer for evaluating large language models in social-science research". *Advances in Methods and Practices in Psychological Science* 8 (2). https://doi.org/10.1177/25152459251325174

[2] Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. 2020. "Language models are few-shot learners". *Advances in neural information processing systems* 33: 1877-1901.

[3] Ciletti, Michele. 2025. "Prompting the muse: Generating prosodically-correct Latin speech with large language models." In *Proceedings of the 63rd Annual*

2 https://lila-erc.eu/1st-lila-ws/. See also the related paper in this volume.

3 https://www.mqdq.it/.

Meeting of the Association for Computational Linguistics 4: Student Research Workshop, edited by Jin Zhao, Mingyang Wang, and Zhu Liu, 740-745.

[4] Ciletti, Michele. 2025. "Veras audire et reddere voces: A corpus of prosodically-correct latin poetic audio from large-language-model tts". In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*.

[5] Comanici, Gheorghe, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein et al. 2025. "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilitie". arXiv preprint arXiv:2507.06261.

[6] Fortson IV, Benjamin W. 2011. "Latin prosody and metrics". In *A companion to the Latin language*: 92-104. Blackwell Publishing Ltd. https://doi.org/10.1002/9781444343397.ch7.

[7] Gil, Alex, and Élika Ortega. 2016. "Global outlooks in digital humanities: Multilingual practices and minimal computing". In *Doing digital humanities*, 58-70. Routledge.

[8] Hurst, Aaron, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, A. J. Ostrow et al. 2024. "Gpt-4o system card". arXiv preprint arXiv:2410.21276.

[9] Johnson, Kyle P., 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,* ed. by Heng Ji, Jong C. Park, Rui Xia, 20-29. *Association for Computational Linguistics*.

[10] Lam, Perry, Huayun Zhang, Nancy F. Chen, Berrak Sisman, and Dorien Herremans. "PRESENT: Zero-Shot Text-to-Prosody Control". *IEEE Signal Processing Letters* 32: 776 - 780. https://doi.org/10.1109/LSP.2025.3528359.

[11] Marvin, Ggaliwango, Nakayiza Hellen, Daudi Jjingo, and Joyce Nakatumba-Nabende. "Prompt engineering in large language models". 2023. In *International conference on data intelligence and cognitive informatics*, 387-402. Springer Nature Singapore.

[12] Mosqueira-Rey, Eduardo, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. 2023. "Human-in-the-loop machine learning: a state of the art." *Artificial Intelligence Review* 56 (4): 3005-3054.

[13] Passarotti, Marco, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. 2020. "Interlinking through lemmas. the lexical collection of the lila

knowledge base of linguistic resources for latin". *Studi e Saggi Linguistici* 58 (1): 177-212.

[14] Reddy, G. Pradeep, YV Pavan Kumar, and K. Purna Prakash. "Hallucinations in large language models (LLMs)". 2024. In *2024 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream)*, 1-6. IEEE.