AI-Driven Approaches to Latin Language Modeling

Data Driven Humanities Team*

Background

The **Data Driven Humanities Research Group** is at the forefront of integrating artificial intelligence with classical philology. This poster presents the latest work of our group, showcasing our research in AI and machine learning for Latin language modeling. Our research focuses on developing a Latin language model by integrating lemmatization, sentence modeling, and prompt engineering, both through traditional querying and automated methods, to enhance linguistic processing, textual comprehension, and language learning.

Latin Lemmatizer

We developed a Latin Lemmatizer that displays declension tables for any queried word(s). The basis of the lemmatizer comes from an online Latin dictionary website. For each word, webscraping is utilized to access every single table and its corresponding word forms. There are functions which detect the part of speech of a given word, the declensions of a table, and printing out all possible words that the website has access to. Right now, all parts of speech work (i.e., noun, verb, adjective, adverb, pronoun, preposition, conjunction, interjection), though there are some minor caveats: words with multiple possible origins (e.g., pulla). The algorithm right now throws an error when this happens, and it asks the user to specify the part of speech they are looking for. If the user re-queries with the desired word and part of speech, then it will choose the first result with that part of speech (see figure 1). Subsequently, we use the lemmatizer for sentence generation to pretrain our sentiment analysis model.

Latin Sentiment Analysis Model

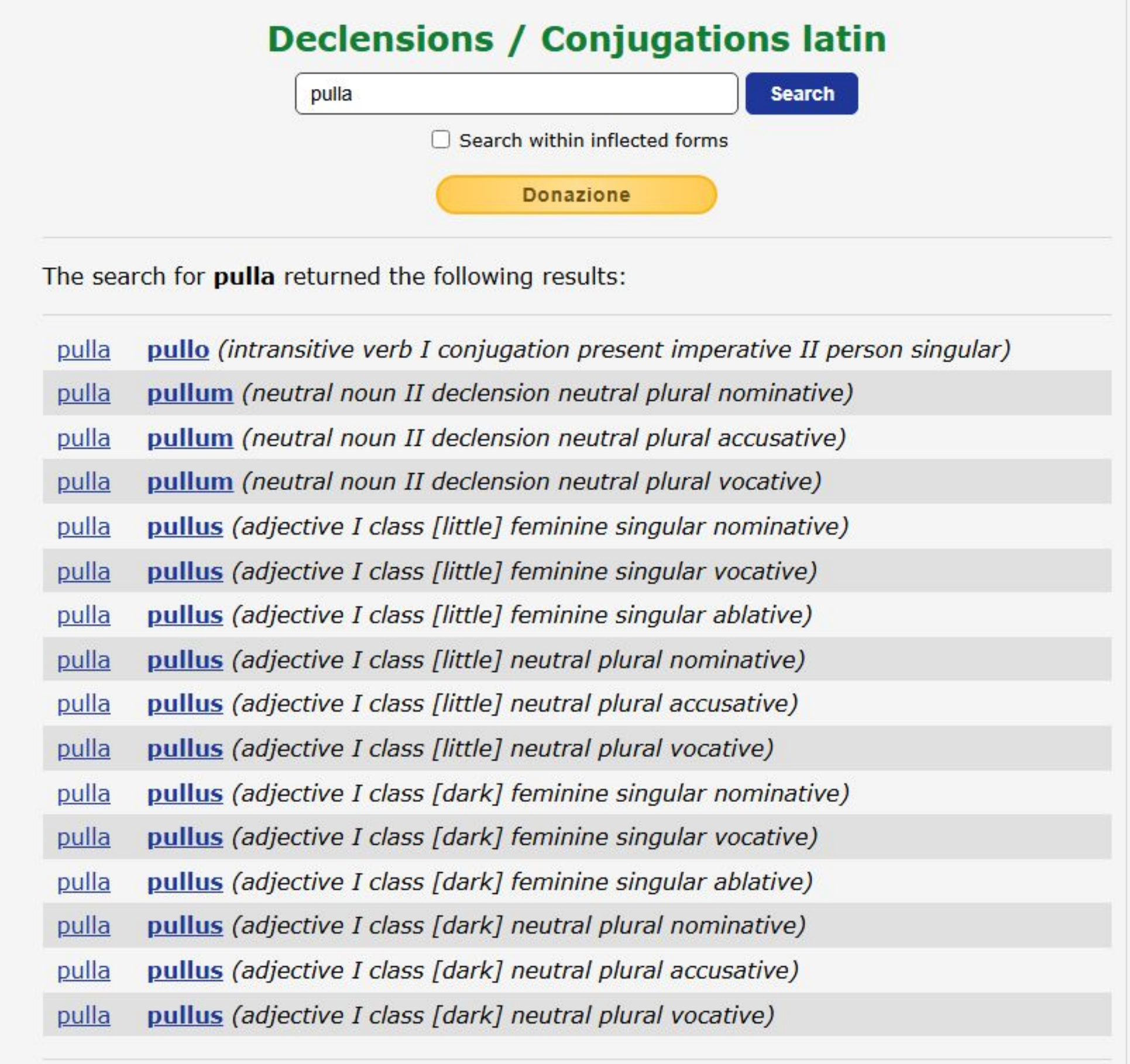
Our Latin Sentiment Analysis system is built upon the *"pnadel/LatinBERT"* model, which is a specialized BERT model pre-trained on Latin text. We modified this base model for sentiment analysis by:

1. Replacing the original masked language modeling head with a sequence classification head
2. Configuring the model for binary classification (positive/negative sentiment)
3. Implementing a custom tokenizer (*"pnadel/latin_tokenizer"*) specifically designed for Latin text
4. Adding padding token configuration to handle variable-length sentences

PNADel/LatinBERT

PNADel/LatinBERT (Hugging Face Model)

- **Developer:** Hugging Face user/org `PNADel`
- **Platform:** Fully trained and deployed via Hugging Face Transformers
- **Architecture: BERT-base** or similar, pretrained from scratch (not based on mBERT)
- **Corpus:** Likely a **larger, more heterogeneous corpus**, possibly including Classical, Medieval, and Ecclesiastical Latin
- **Purpose:** Broader applicability — usable for a wide variety of **downstream NLP tasks** in Latin via transfer learning

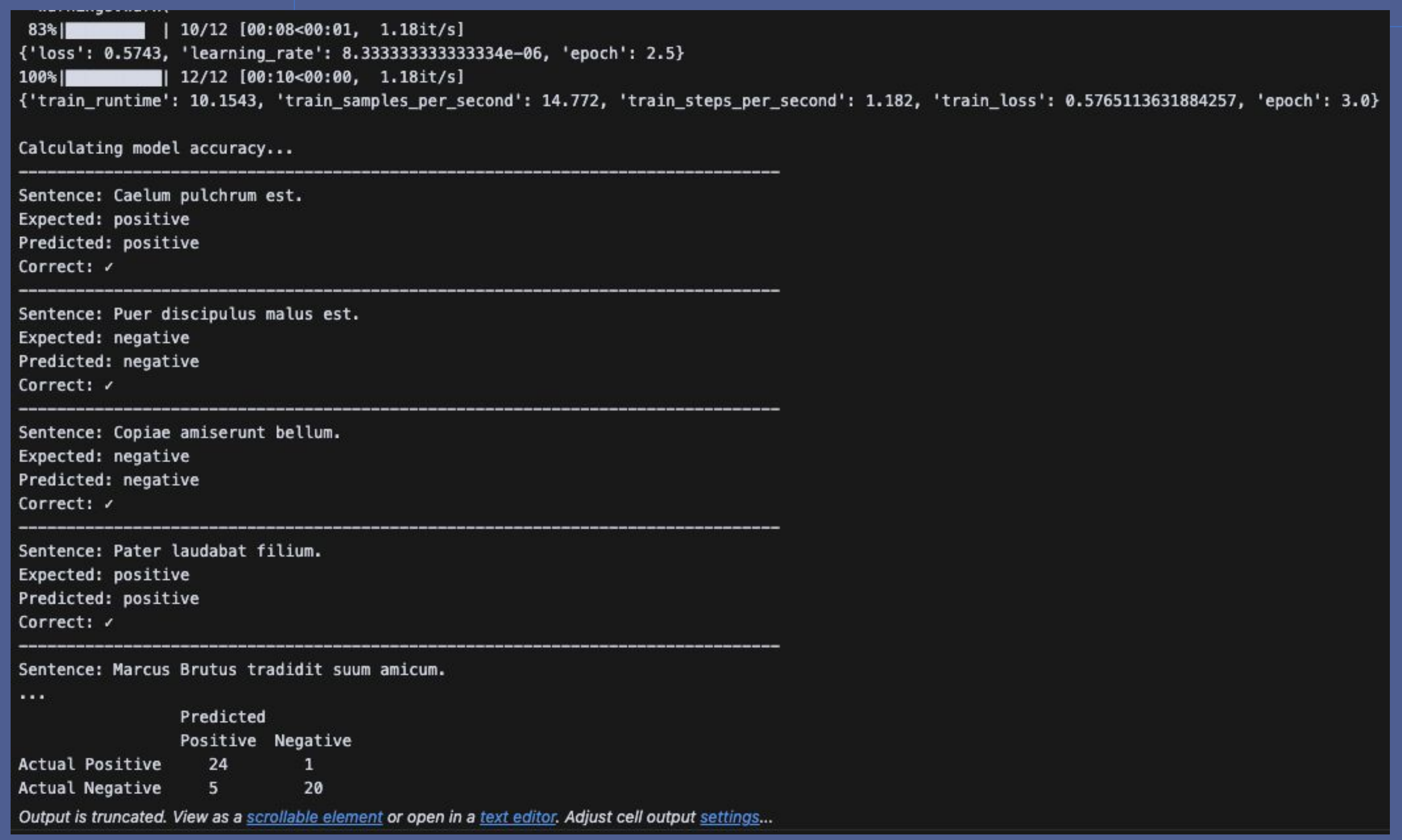


Declensions / Conjugations latin

The search for **pulla** returned the following results:

pulla	**pullo** *(intransitive verb I conjugation present imperative II person singular)*
pulla	**pullum** *(neutral noun II declension neutral plural nominative)*
pulla	**pullum** *(neutral noun II declension neutral plural accusative)*
pulla	**pullum** *(neutral noun II declension neutral plural vocative)*
pulla	**pullus** *(adjective I class [little] feminine singular nominative)*
pulla	**pullus** *(adjective I class [little] feminine singular vocative)*
pulla	**pullus** *(adjective I class [little] feminine singular ablative)*
pulla	**pullus** *(adjective I class [little] neutral plural nominative)*
pulla	**pullus** *(adjective I class [little] neutral plural accusative)*
pulla	**pullus** *(adjective I class [little] neutral plural vocative)*
pulla	**pullus** *(adjective I class [dark] feminine singular nominative)*
pulla	**pullus** *(adjective I class [dark] feminine singular vocative)*
pulla	**pullus** *(adjective I class [dark] feminine singular ablative)*
pulla	**pullus** *(adjective I class [dark] neutral plural nominative)*
pulla	**pullus** *(adjective I class [dark] neutral plural accusative)*
pulla	**pullus** *(adjective I class [dark] neutral plural vocative)*

Figure 1: Web response when querying "pulla"



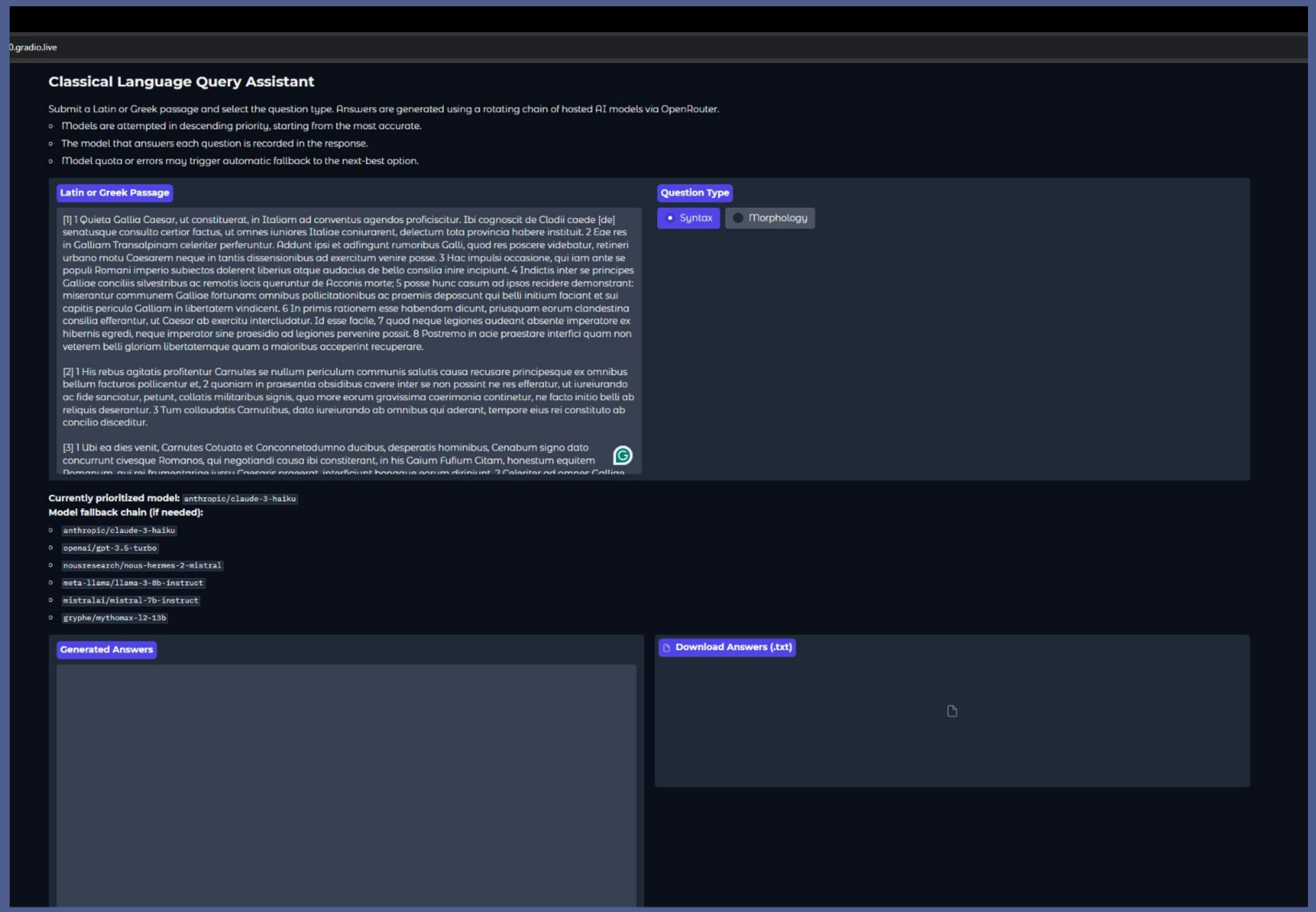
```
 83%|████████  | 10/12 [00:08<00:01,  1.18it/s]
{'loss': 0.5743, 'learning_rate': 8.333333333333334e-06, 'epoch': 2.5}
100%|██████████| 12/12 [00:10<00:00,  1.18it/s]
{'train_runtime': 10.1543, 'train_samples_per_second': 14.772, 'train_steps_per_second': 1.182, 'train_loss': 0.5765113631884257, 'epoch': 3.0}

Calculating model accuracy...
--------------------------------------------------------------------------------
Sentence: Caelum pulchrum est.
Expected: positive
Predicted: positive
Correct: ✓
--------------------------------------------------------------------------------
Sentence: Puer discipulus malus est.
Expected: negative
Predicted: negative
Correct: ✓
--------------------------------------------------------------------------------
Sentence: Copiae amiserunt bellum.
Expected: negative
Predicted: negative
Correct: ✓
--------------------------------------------------------------------------------
Sentence: Pater laudabat filium.
Expected: positive
Predicted: positive
Correct: ✓
--------------------------------------------------------------------------------
Sentence: Marcus Brutus tradidit suum amicum.
...
                 Predicted
                 Positive  Negative
Actual Positive     24        1
Actual Negative     5         20
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...
```

Figure 2. Our modified model achieved an impressive 88% accuracy on the test set, demonstrating strong performance in classifying Latin sentences. The model successfully handles various Latin sentence structures and maintains consistent performance across different types of conten.

Query Engineering and Automation

Another focus of our project is the development and use of innovative methods in computational philology through **advanced query engineering** for Greek and Latin texts. Initially, our researchers manually crafted complex queries using ChatGPT-4 to analyze syntactic structures in ancient languages, allowing for precise linguistic and interpretive insights.
Building on this foundational work, the team has now programmed these processes into an automated system, training the model to recognize and process Greek and Latin syntax autonomously.



Classical Language Query Assistant

Submit a Latin or Greek passage and select the question type. Answers are generated using a rotating chain of hosted AI models via OpenRouter.

- Models are attempted in descending priority, starting from the most accurate.
- The model that answers each question is recorded in the response.
- Model quota or errors may trigger automatic fallback to the next-best option.

Latin or Greek Passage

[1] 1 Quieta Gallia Caesar, ut constituerat, in Italiam ad conventus agendos proficiscitur. Ibi cognoscit de Clodii caede [de] senatusque consulto certior factus, ut omnes iuniores Italiae coniurarent, delectum tota provincia habere instituit. 2 Eae res in Galliam Transalpinam celeriter perferuntur. Addunt ipsi et adfingunt rumoribus Galli, quod res poscere videbatur, retineri urbano motu Caesarem neque in tantis dissensionibus ad exercitum venire posse. 3 Hac impulsi occasione, qui iam ante se populi Romani imperio subiectos dolerent liberius atque audacius de bello consilia inire incipiunt. 4 Indictis inter se principes Galliae conciliis silvestribus ac remotis locis queruntur de Acconis morte; 5 posse hunc casum ad ipsos recidere demonstrant: miserantur communem Galliae fortunam: omnibus pollicitationibus ac praemiis deposcunt qui belli initium faciant et sui capitis periculo Galliam in libertatem vindicent. 6 In primis rationem esse habendam dicunt, priusquam eorum clandestina consilia efferantur, ut Caesar ab exercitu intercludatur. Id esse facile, 7 quod neque legiones audeant absente imperatore ex hibernis egredi, neque imperator sine praesidio ad legiones pervenire possit. 8 Postremo in acie praestare interfici quam non veterem belli gloriam libertatemque quam a maioribus acceperint recuperare.

[2] 1 His rebus agitatis profitentur Carnutes se nullum periculum communis salutis causa recusare principesque ex omnibus bellum facturos pollicentur et, 2 quoniam in praesentia obsidibus cavere inter se non possint ne res efferatur, ut iureiurando ac fide sanciatur, petunt, collatis militaribus signis, quo more eorum gravissima caerimonia continetur, ne facto initio belli ab reliquis deserantur. 3 Tum collaudatis Carnutibus, dato iureiurando ab omnibus qui aderant, tempore eius rei constituto ab concilio disceditur.

[3] 1 Ubi ea dies venit, Carnutes Cotuato et Conconnetodumno ducibus, desperatis hominibus, Cenabum signo dato concurrunt civesque Romanos, qui negotiandi causa ibi constiterant, in his Gaium Fufium Citam, honestum equitem

Question Type

● Syntax ○ Morphology

Currently prioritized model: `anthropic/claude-3-haiku`
Model fallback chain (if needed):

- `anthropic/claude-3-haiku`
- `openai/gpt-3.5-turbo`
- `nousresearch/nous-hermes-2-mistral`
- `meta-llama/llama-3-8b-instruct`
- `mistralai/mistral-7b-instruct`
- `gryphe/mythomax-l2-13b`

Figure 3: Query system interface

Future Directions

Future directions include the further development of both the Greek and Latin lemmatizers, sentence modeling, and sentiment analysis models.

Furthermore, we are exploring the development of our own chatbot designed to interact with users in Ancient Greek and Latin. This initiative aims to bridge the gap between modern AI technologies and classical studies, fostering engagement with ancient languages in innovative ways.

***Data Driven Humanities Team Members:**
Justine Asman, Wavid Bowman, Eleni Bozia, Aidan Burrowes, Thomas Cerniglia, Srija Dey, Gebril Fradj, Nicole Fong, Guhan Gnanam, Zach Hracho, Jacob Hoppenstedt, Jeffrey James, Eden Layman, Jonnhy Liu, Daniel Miller, Connor Munjed, Niketha Nethaji, Krish Patel, Gillian Rodgers, Shae Robinson, Trey Slaten, Abraham Stefanos, Jordan Yu