

# Interlinea: Validated Language-Model Annotation and Rules-First Scansion of Latin and Ancient Greek

Rumi Elías Calles\*

Aeterna Institute · October 2026

**Abstract.** Interlinea annotates every word of a Latin or Ancient Greek text with its lemma, part of speech, morphology, syntactic head and a gloss, and scans its verse. A language model does the analysis, but code owns the form of the answer. Code numbers the words, the model fills a Universal Dependencies schema for each number, a validator checks the tree and the features, and a failed check returns to the model once. For verse, rules enumerate the readings the metre allows and the model only picks one. I evaluate the design on 833 sentences of expert annotation, including two sets released after the training cutoffs of all four models tested, against Stanza and UDPipe 2 trained on each test treebank. Claude Opus 5.5 finds heads better than either parser, by 5.5 UAS points over UDPipe 2 in Latin and 8.5 in Greek, while the parsers keep the lead on exact morphology and relation labels. Numbered words are the protocol's most valuable part, and without them a model loses 15.6 UAS points. The correction turn makes nine in ten malformed answers valid without changing their accuracy. In scansion, a model choosing among the rules' candidates is right on 78.6% of the lines the rules leave open, against 43.1% when it scans whole lines.

**Keywords:** Latin; Ancient Greek; large language models; Universal Dependencies; structured output; automatic scansion

## 1 INTRODUCTION

Γραμματική ἐστὶν ἐμπειρία τῶν παρὰ ποιηταῖς τε καὶ  
συγγραφεῦσιν ὥς ἐπὶ τὸ πολὺ λεγομένων.  
Grammar is the practical knowledge of what poets and  
prose writers usually say.  
Dionysius Thrax, *Ars Grammatica* 1

A reader of Latin or Greek spends most of an hour on the forms of words. In Figure 1, *Lesbia* could be nominative, vocative or ablative, and *amemus* is a subjunctive that a beginner reads as a future. Morphological tools such as Morpheus (Crane, 1991) and Alpheios (Alpheios) list every analysis a form allows and leave the reader to choose.

Large language models can make that choice, and with a few examples they match trained taggers on lemmas and parts of speech (Vidal-Gorène et al., 2026). They also fail in ways a learner cannot see. Asked for twelve words, a model returns eleven, merges an enclitic, gives a form an impossible case, or builds a tree with two roots.

---

\*Correspondence: [aeterna-institute.org/contact](https://aeterna-institute.org/contact).

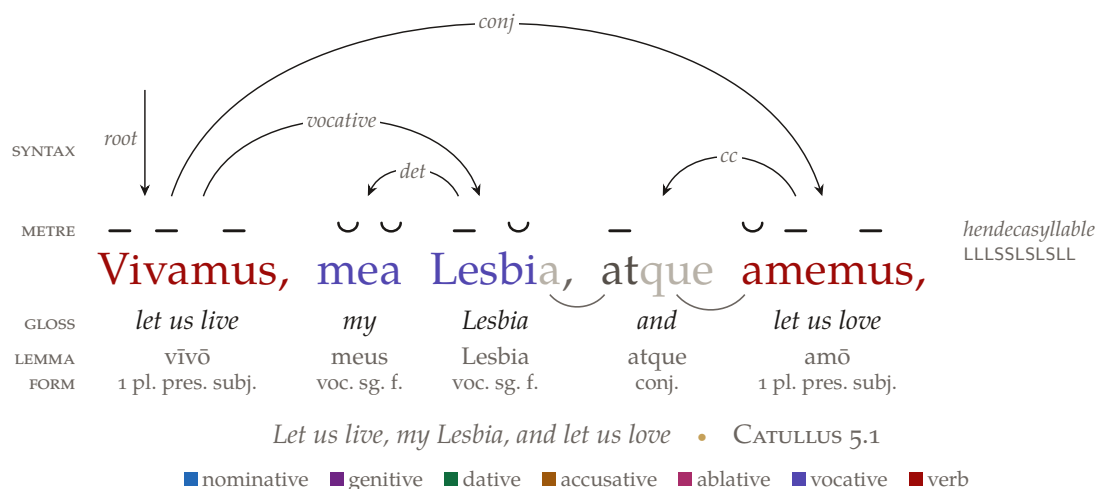


FIGURE 1. The opening of Catullus 5 as Interlinea shows it: the stored dependency tree and the hendecasyllable's quantities above the words, case and verb colours on them, and a gloss, the lemma and the form below. A model wrote the analysis and the validator of §3.2 accepted it. The rules of §3.4 left four scansiones and two models agreed on one.

Interlinea is a reading app built on one rule. Code owns the form of the answer, and the model only chooses. Code numbers the words and asks for one object per number, so the model never writes the text. A validator checks the tree and the features, and a failed check goes back to the model once. For verse, rules enumerate the scansiones the metre allows and a model picks one by number. This paper makes three contributions.

- An annotation protocol in which the model never writes the tokens it annotates, with a formal validity predicate and one correction turn (§3).
- Rules-first scansion, in which prosody narrows a line to its metrically valid readings and models choose among them (§3.4).
- An evaluation on 833 sentences of expert annotation, two sets of them released after every tested model's training cutoff, against parsers trained on each test treebank, with ablations of each part of the protocol (§4–6).

Claude Opus 5.5 finds syntactic heads better than either parser, by 5.5 UAS points over UDPipe 2 in Latin and 8.5 in Greek, while the parsers keep exact morphology and relation labels. Numbered words are worth 15.6 UAS points to a model. The correction turn makes nine in ten malformed answers valid without changing their accuracy. And models choosing among candidate scansiones are far more accurate than models scanning whole lines.

## 2 RELATED WORK

**Treebanks and parsers.** The Ancient Greek and Latin Dependency Treebanks (Bamman and Crane, 2011) and PROIEL (Haug and Jøhndal, 2008) annotate classical and biblical texts, and Universal Dependencies (Nivre et al., 2020; Zeman et al., 2026) brought them into one scheme with newer treebanks such as the Index Thomisticus (Passarotti, 2019), the Lucca charters of LLCT (Cecchini et al., 2020a), UDante (Cecchini et al., 2020b), CIRCSE (Iurescia et al., 2024) and the Septuagint of PTNK (Swanson et al., 2024). GLAUx annotates Greek automatically at the scale of a corpus (Keersmaekers, 2021), and LEMLAT analyses Latin forms from a lexicon (Passarotti et al., 2017). Parsers trained on these treebanks do well in their own domain and lose much of it outside. Stanza (Qi et al., 2020) parses LLCT at 96 UAS and Perseus Latin at 72, and the EvaLatin campaigns have measured the same drop across genres (Iurescia et al., 2026; Sprugnoli et al., 2020, 2022, 2024). The strongest systems there build on pretrained encoders, as LatinPipe does (Straka et al., 2024) with Latin BERT (Bamman and Burns, 2020) among others. I compare with Stanza and UDPipe 2 (Straka, 2018), each trained on the test set’s own treebank.

**Language models as annotators.** Pretrained models of Latin and Greek help tagging and lemmatization (Riemenschneider and Frank, 2023). Prompted general models have tagged sixteenth-century Latin (Stüssi and Ströbel, 2024), and few-shot GPT-4 and Mistral match a trained tagger on lemmas and parts of speech for Ancient Greek, Classical Armenian, Old Georgian and Syriac (Vidal-Gorène et al., 2026). Fine-tuned encoders remain strong for unedited Byzantine Greek (Swaelens et al., 2025), and a fine-tuned LLaMA can produce UD trees as text (Hromei et al., 2024). The closest work to mine is THIVLVC, in which a model revises UDPipe’s parse with retrieved treebank sentences and gains 17 CLAS points on Seneca’s verse (Pommeret et al., 2026). In Interlinea the model writes the whole analysis, including the gloss a learner needs, and code decides whether it may stand.

**Structured output and correction.** Constrained decoding makes output match a grammar or schema by construction (Geng et al., 2023; Willard and Louf, 2023), though strict formats can cost reasoning accuracy (Tam et al., 2024). Libraries that validate a model’s output and retry on failure are common in practice (Liu and Leo, 2026). Self-correction without outside feedback does not reliably help (Huang et al., 2024), and a survey finds it works when the feedback is external and reliable (Kamoi et al., 2024), as with tool critiques (Gou et al., 2024) or execution results (Chen et al., 2024b), rather than the model’s own review (Madaan et al., 2023; Shinn et al., 2023). Interlinea’s correction turn is of the external kind, since a deterministic validator writes the feedback. Its scansion uses agreement between two models from different providers, where self-consistency samples one model several times (Wang et al., 2023). Its fallback order resembles a cascade (Chen et al., 2024a), ordered for speed rather than price.

**Contamination.** Public test sets may be in a model’s training data (Sainz et al., 2023), and accuracy on data released after a model’s cutoff is a direct check (Roberts et al., 2024). Two of my gold sets were released after all four cutoffs.

**Scansion.** Latin and Greek verse is scanned from syllable quantities that follow from position, from nature or from the metre itself (Raven, 1965; West, 1982). Finite-state scanners for the hexameter work this way (CLTK; Schumann et al., 2021), and large scanned corpora exist (Chamberlain, 2026; Colombi et al., 2011). Recent work learns vowel length from data (Cullhed and Thörn Cleland, 2026; Thörn Cleland and Cullhed, 2026) or prompts models to scan whole lines (Ruiz Fabo et al., 2026). Interlinea keeps the rules and asks a model only what they leave open.

### 3 METHOD

Figure 2 shows what happens to a sentence the first time anyone opens it, and Figure 3 what the reader sees. In production, GPT-6 luna annotates every sentence a reader opens, with Gemini 3.8 Flash as its fallback. Each analysis is stored under the SHA-256 hash of its language, schema version and NFC-normalized text, so a library sentence is paid for once and served to every later reader.

#### 3.1 Index-only annotation

The tokenizer splits a sentence into words  $w_1, \dots, w_n$ . It keeps the Latin enclitics *-que*, *-ne* and *-ve* on their host, as a reader sees them, and keeps punctuation out. The model receives the sentence and the numbered list and returns one object per index,

$$a_i = (i, \ell_i, e_i, u_i, f_i, h_i, r_i, g_i, \nu_i, c_i),$$

with lemma  $\ell_i$ , dictionary entry  $e_i$ , UD part of speech  $u_i$ , UD features  $f_i$ , head  $h_i \in \{0, \dots, n\}$ , UD relation  $r_i$ , English gloss  $g_i$ , a note  $\nu_i$  and an enclitic  $c_i$ . The model never writes  $w_i$ , so it cannot drop, merge or misspell a word. Heads are indices, so a tree may

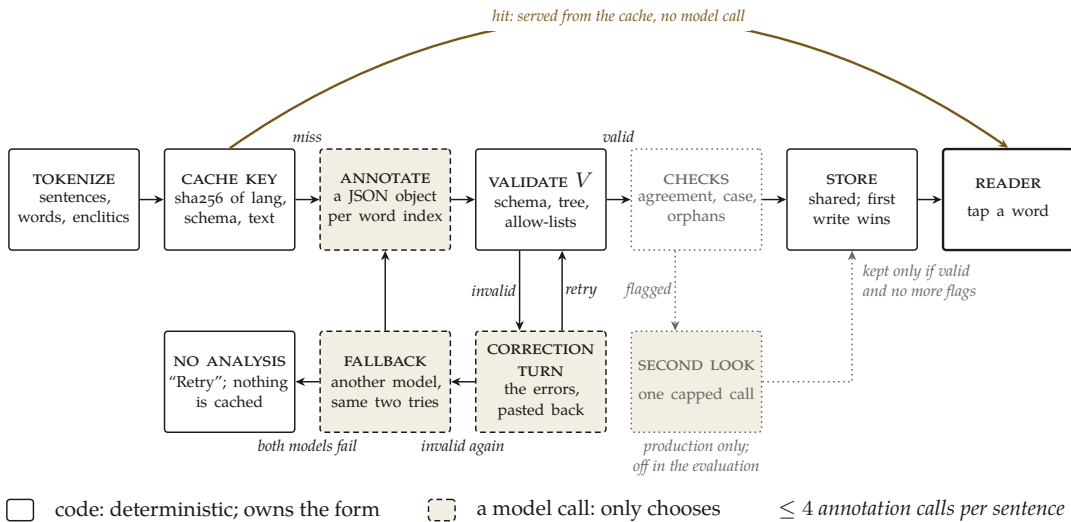


FIGURE 2. One sentence’s path. A cached sentence goes straight to the reader. Otherwise the model annotates numbered words, and the validator  $V$  passes the answer or returns its errors for one correction turn. A second failure moves to a fallback model with the same two tries, so at most four calls are made, and a failure is never cached. Dotted boxes run in production only and were off in the evaluation.

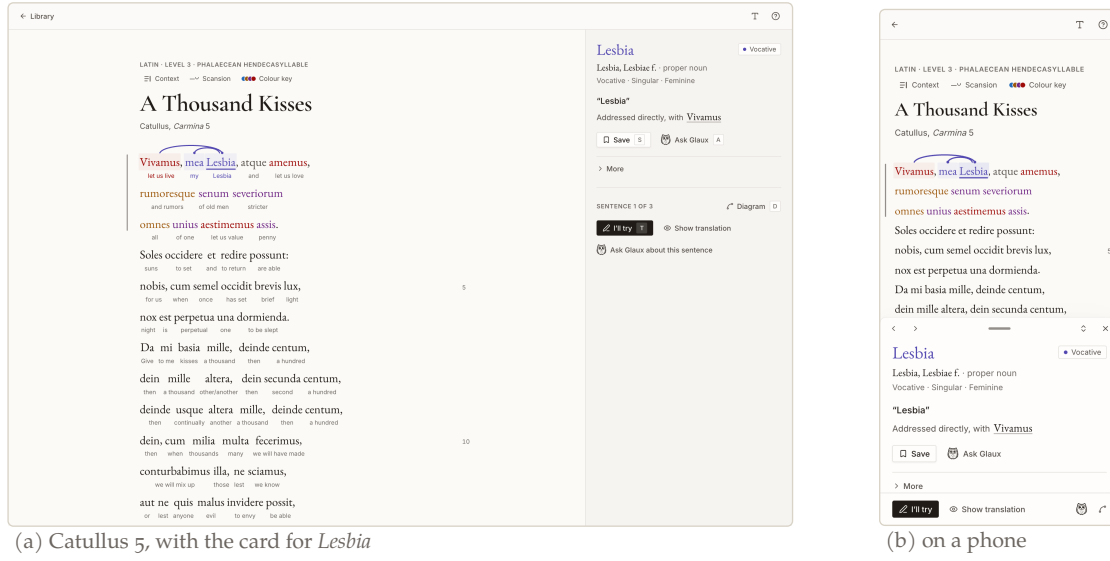


FIGURE 3. The reader. (a) Catullus 5 with interlinear glosses and the card for *Lesbia*. (b) The same card on a phone.

cross itself or a line break (Figure 4). The schema is JSON Schema in the subset every provider’s strict structured output accepts, and the prompt states the UD conventions and gives two worked sentences per language (Appendix A).

### 3.2 The validity predicate

**DEFINITION** *Valid annotation.* An annotation  $A$  of a sentence with  $n$  words in language  $L$  is valid,  $V(A) = 1$ , when all of the following hold.

- (i)  $A$  matches the schema, and  $|A| = n$  with  $a_i.i = i$  for every  $i$ .
- (ii)  $h_i \in \{0, \dots, n\}$  and  $h_i \neq i$  for every  $i$ .
- (iii) Exactly one  $i$  has  $h_i = 0$ , and that word, and only it, has  $r_i = \text{root}$ .
- (iv) The graph with an edge  $h_i \rightarrow i$  for every  $h_i > 0$  has no cycle, so with (iii) it is a tree.
- (v)  $\ell_i, g_i$  and the translation are not empty, and every feature in  $f_i$  is in the allow-list of  $L$ .
- (vi) For Greek, a note that names the aorist goes with Tense=Past | Aspect=Perf.

Each failed condition becomes a line of plain English the model can act on, such as token 4: Mood=Part is not allowed (use Ind, Sub, Imp).

### 3.3 The correction turn

An invalid answer gets one correction turn, in which the model sees its answer and the validator’s messages and answers again. If that fails, a fallback model gets the same two tries. Calls run at minimal reasoning with a 16,000-token output cap. In production, a second model may also review a valid answer that grammar rules flag, such as an adjective that agrees with no noun. That second look was off in the evaluation.

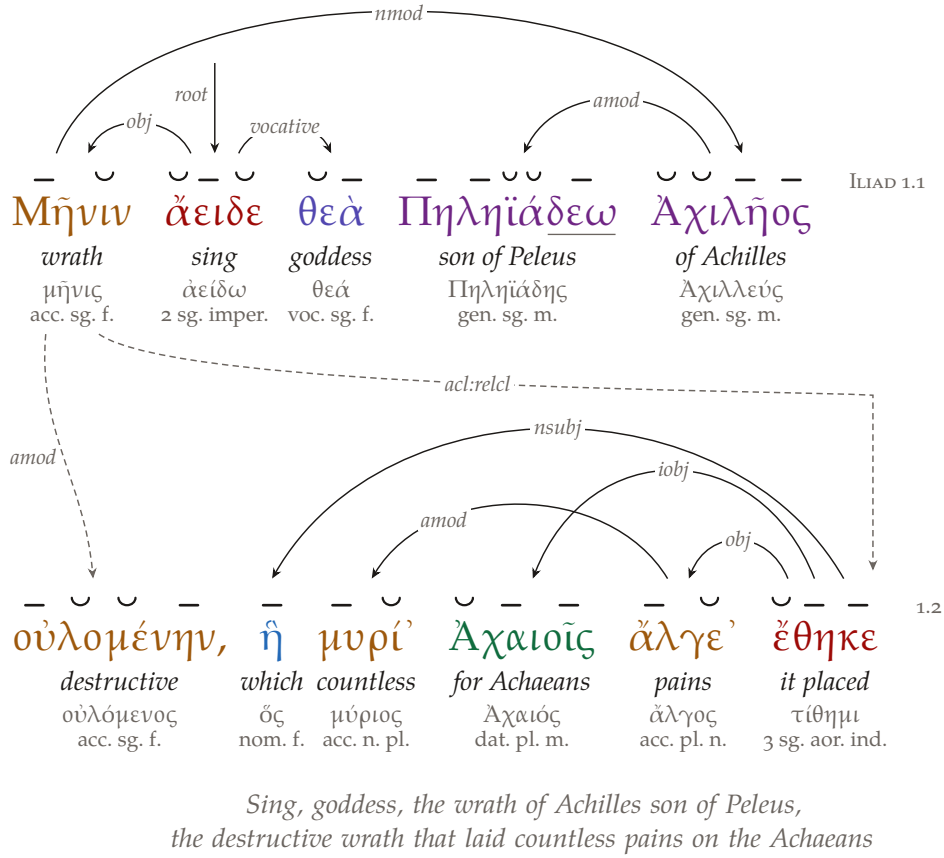


FIGURE 4. The first two lines of the *Iliad*. Heads are word indices, so the tree crosses the line break, where οὐλομένην and the relative clause both depend on μῆνιν, and crosses itself around the hyperbaton μυρί' Ἀχαιοῖς ἄλγε'. The rules scan both lines alone.

### 3.4 Rules-first scansion

Rules split a line into syllables and mark each quantity long, short or unknown, from position, nature, elision and synizesis. Let  $q_j \in \{L, S, ?\}$  be the rules' verdict on syllable  $j$ . For a metre  $M$ , a regular language over  $\{L, S\}$ , the candidates are

$$C(M) = \{ \pi \in \{L, S\}^m : \pi \in M, \pi_j = q_j \text{ wherever } q_j \neq ? \}.$$

With one candidate the line is scanned without a model. With two to eight, a model sees the numbered candidates and returns a number. Otherwise the line stays plain. In dactylic metres one model's pick is kept. Elsewhere two models from different providers must agree, or the line stays plain, since a plain line costs a reader less than a wrong one (Figure 5).

|            |     |    |     |    |   |     |    |     |    |       |   |    |     |                                                                              |
|------------|-----|----|-----|----|---|-----|----|-----|----|-------|---|----|-----|------------------------------------------------------------------------------|
| LINE       | Vi  | va | mus | me | a | Les | bi | (a) | at | (que) | a | me | mus |                                                                              |
| RULES      | ?   | ?  | —   | ∪  | ? | —   | ∪  | —   | ?  | ?     | ? |    |     | <i>vī, vā</i> : long by nature,<br>invisible to the rules<br>without macrons |
| METRE      | ×   | ×  | —   | ∪  | ∪ | —   | ∪  | —   | ∪  | —     | × |    |     |                                                                              |
| CANDIDATES | (1) | ∪  | ∪   | —  | ∪ | ∪   | —  | ∪   | —  | ∪     | — | —  |     | MODEL A: (4)                                                                 |
|            | (2) | —  | ∪   | —  | ∪ | ∪   | —  | ∪   | —  | ∪     | — | —  |     | MODEL B: (4)                                                                 |
|            | (3) | ∪  | —   | —  | ∪ | ∪   | —  | ∪   | —  | ∪     | — | —  |     | they agree, so                                                               |
|            | (4) | —  | —   | —  | ∪ | ∪   | —  | ∪   | —  | ∪     | — | —  |     | the line is stored                                                           |

FIGURE 5. The rules leave a model one question. In Catullus 5.1, position and the hendecasyllable fix every syllable but the free opening pair, so four candidates remain, and they differ only where knowing that *vīvāmus* has two long vowels decides.

## 4 EXPERIMENTAL SETUP

**Gold data.** I used existing expert annotation only (Table 1). For Latin, the test splits of UD CIRCSE (Seneca and Tacitus) and UD LLCT (Lucca charters), and the test sentences of UD Perseus. For Greek, the test sentences of UD Perseus and the test split of UD PTNK (the Septuagint). UD’s Perseus conversion is non-commercial, so for those sentences I took the annotation from the Ancient Greek and Latin Dependency Treebank (Bamman and Crane, 2011) and converted it to UD myself. Two more sets, the DALME-Marseille inventories of 1258 to 1446 (Pizzorno, 2026) and the UNESP treebank of Artemidorus (Ferreira, 2021), were released in August 2026. The models’ training data end in June 2026 for Claude Opus 5.5 (Anthropic, 2026), May 2026 for GPT-6 luna (OpenAI, 2026) and March 2026 for Gemini 3.8 Flash (Google DeepMind, 2026). A sentence is used only if Interlinea’s tokenizer reproduces its gold words exactly.

**Systems.** Four models ran through Interlinea’s annotation function with the production prompt, minimal reasoning and one correction turn, with no fallback and no second look: Claude Opus 5.5, Gemini 3.8 Flash, GPT-6 luna and Gemini 2.5 Flash. The baselines are Stanza 1.15 (Qi et al., 2020) and UDPipe 2 (Straka, 2018), given the gold words, each with the model trained on the test set’s own treebank, chosen in advance (LLCT for the inventories and Perseus for Artemidorus). Every model was called through the Vercel AI Gateway, and prices are its list prices on 19 October 2026.

**Metrics.** Lemma accuracy ignores diacritics, case and *u/v*, *i/j* spelling. Features are scored one at a time for case, number, gender, tense, mood, voice and person. UAS counts correct heads, and LAS, on UD-native sets only, also the relation. A sentence with no valid answer scores zero. The main scores accept both sides of the conventions on which treebanks disagree, such as PRON or DET for a demonstrative (Appendix B); §5.2 reports strict scores.

**Statistics.** Systems are compared with a paired bootstrap over sentences, stratified by gold set, with 10,000 resamples and 95% percentile intervals (Berg-Kirkpatrick et al., 2012; Dror et al., 2018; Efron and Tibshirani, 1993). P-values are Holm-adjusted (Holm, 1979) within three families: models against parsers (64 tests), ablations, and models against

| Set                                         | Texts                          | Public  | Sent. | Words | Licence         |
|---------------------------------------------|--------------------------------|---------|-------|-------|-----------------|
| PUBLIC TEST SPLITS: MAIN RESULTS            |                                |         |       |       |                 |
| UD Latin CIRCSE (test)                      | Seneca,<br>tragedy;<br>Tacitus | 2024    | 100   | 1,135 | CC BY-SA 4.0    |
| UD Latin LLCT (test)                        | Lucca charters,<br>8th–9th c.  | 2020    | 75    | 1,043 | CC BY-SA 4.0    |
| AGDT Latin (UD Perseus test)                | classical prose<br>and verse   | 2011    | 75    | 669   | CC BY-SA 3.0 US |
| AGDT Greek (UD Perseus test)                | Homer to<br>Plutarch           | 2011    | 170   | 2,450 | CC BY-SA 3.0 US |
| UD Greek PTNK (test)                        | Septuagint                     | 2023    | 80    | 1,418 | CC BY-SA 4.0    |
| RELEASED AFTER THE MODELS’ TRAINING CUTOFFS |                                |         |       |       |                 |
| DALME-Marseille gold                        | inventories,<br>1258–1446      | 2026-08 | 193   | 1,160 | CC BY 4.0       |
| UNESP Artemidorus                           | prose, 2nd c.<br>CE            | 2026-08 | 140   | 2,146 | CC BY-SA 4.0    |

TABLE 1. Gold sets. The public sets give the main results, 250 sentences per language drawn with a fixed seed from the UD v2.18 test splits. The two later sets are used whole.

GPT-6 luna.

## 5 RESULTS

|                    |                            | word accuracy (%) |       |      |          |      | per sentence     |       |   |
|--------------------|----------------------------|-------------------|-------|------|----------|------|------------------|-------|---|
|                    |                            | Valid             | Lemma | UPOS | Features | UAS  | LAS <sup>a</sup> | ¢     | s |
| LATIN              | 250 sentences, 2,847 words |                   |       |      |          |      |                  |       |   |
| Claude Opus 5.5    | 100.0                      | 92.4              | 95.7  | 95.5 | 86.9     | 75.6 | 5.52             | 12.2  |   |
| Gemini 3.8 Flash   | 98.4                       | 89.4              | 92.0  | 92.8 | 83.5     | 72.7 | 4.61             | 102.0 |   |
| GPT-6 luna         | 99.2                       | 88.2              | 93.2  | 92.2 | 75.5     | 63.5 | 0.10             | 10.7  |   |
| Gemini 2.5 Flash   | 99.6                       | 89.8              | 94.2  | 92.6 | 76.6     | 66.2 | 0.51             | 7.1   |   |
| Stanza (matched)   | 100.0                      | 89.3              | 94.2  | 92.2 | 77.8     | 72.8 | ≈0               | <1    |   |
| UDPipe 2 (matched) | 100.0                      | 92.1              | 94.8  | 92.4 | 81.4     | 77.0 | ≈0               | <1    |   |
| GREEK              | 250 sentences, 3,868 words |                   |       |      |          |      |                  |       |   |
| Claude Opus 5.5    | 100.0                      | 94.4              | 94.9  | 97.5 | 91.4     | 87.9 | 6.94             | 23.8  |   |
| Gemini 3.8 Flash   | 97.6                       | 91.8              | 92.2  | 94.5 | 88.5     | 83.1 | 3.84             | 101.6 |   |
| GPT-6 luna         | 100.0                      | 93.7              | 95.1  | 97.0 | 85.8     | 83.2 | 0.13             | 48.7  |   |
| Gemini 2.5 Flash   | 98.8                       | 92.4              | 93.7  | 95.1 | 85.2     | 83.2 | 0.62             | 9.9   |   |
| Stanza (matched)   | 100.0                      | 89.9              | 95.0  | 94.5 | 80.9     | 86.0 | ≈0               | <1    |   |
| UDPipe 2 (matched) | 100.0                      | 90.9              | 95.8  | 94.8 | 82.9     | 88.9 | ≈0               | <1    |   |

TABLE 2. Results on the public test sets. Valid is the share of sentences with a valid answer. <sup>a</sup>LAS on UD-native sets only. Price in cents and wall-clock seconds per sentence.

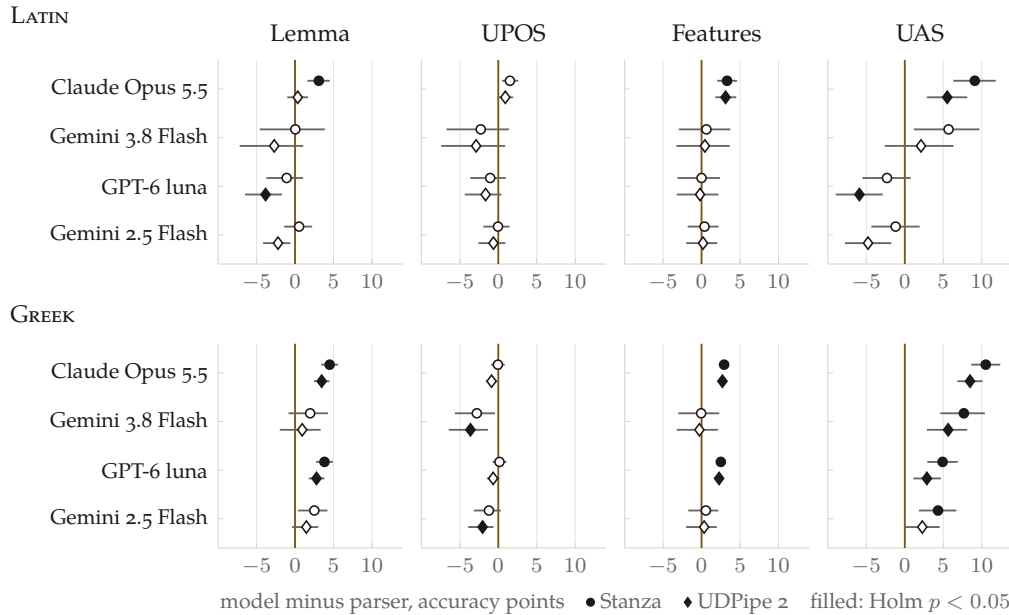


FIGURE 6. The best model finds heads better than both parsers, and words are close. Each point is a model’s accuracy minus a parser’s on the same sentences, with a 95% paired interval, filled where the difference survives Holm correction across the 64 tests.

### 5.1 Models against parsers

Table 2 gives the scores and Figure 6 the paired differences. Claude Opus 5.5 reaches UAS 86.9 in Latin and 91.4 in Greek, 5.5 and 8.5 points above UDPipe 2, the stronger parser. In Greek every model’s UAS is above both parsers’. In Latin GPT-6 luna is 11.4 points behind Opus and below UDPipe 2. Words are closer than trees. Every model gets the seven features right more than nine times in ten, and in Latin Opus and UDPipe 2 are level on lemmas. LAS goes the other way. UDPipe 2 reaches 77.0 in Latin and 88.9 in Greek, against 75.6 and 87.9 for Opus. The models choose heads well and relation labels less well.

### 5.2 Strict scoring and the official trees

Lenient scoring might favour the models, which follow no treebank. Under strict scoring (Table 3a) both parsers lead on morphology: UDPipe 2 gets 87.2% of Latin words exactly right and Opus 81.3%. A parser learns its treebank’s habits, and a model guesses them. My AGDT conversion is the other risk. Against the official UD trees (Table 3b), Opus’s Greek UAS falls from 89.1 to 85.9 and UDPipe 2’s rises from 76.4 to 80.0, so the conversion favoured the models, but Opus still leads.

### 5.3 Each gold set, and text no model has seen

Figure 7 breaks the results down by set. On the Marseille inventories, full of medieval words and names, Opus keeps 82.9 lemma accuracy and UDPipe 2 falls to 67.8. Whatever the models know about Latin, they did not learn it from these sentences. Artemidorus reverses the order, because its trees follow the AGDT guidelines of the parser’s training data and my conversion of them costs the models several UAS points. The two later sets are small and differ in genre, so they cannot rule out contamination of the public results.

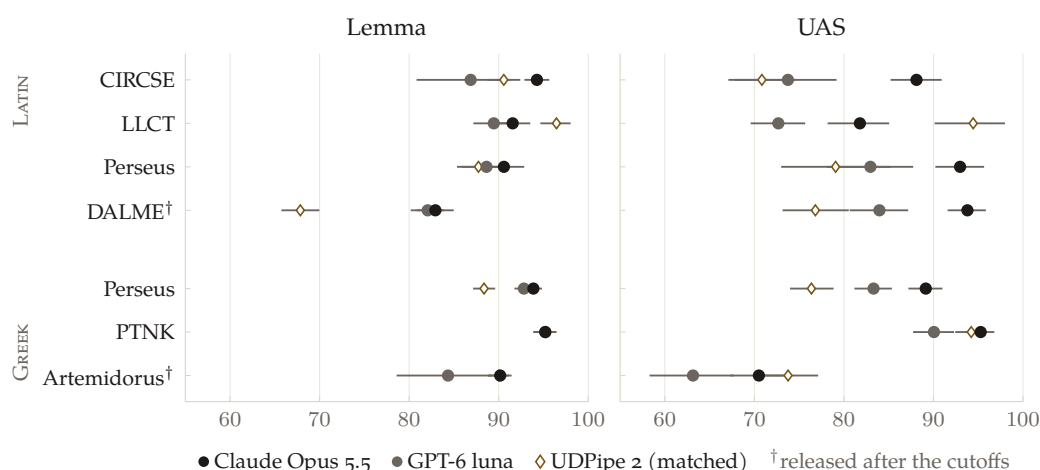


FIGURE 7. The models’ lead over the parser grows with the distance from the parser’s treebank. Accuracy per gold set with 95% intervals. The parser is best on LLCT, its own charters, and on Artemidorus, which follows its annotation guidelines, and worst on the medieval inventories.

(a) Strict scores on the UD-native sets

|                          | UPOS    |        | Feats<br>exact, 7 | LAS  | CLAS |
|--------------------------|---------|--------|-------------------|------|------|
|                          | lenient | strict |                   |      |      |
| LATIN la-circse, la-llct |         |        |                   |      |      |
| Claude Opus 5.5          | 96.4    | 93.5   | 81.3              | 75.6 | 74.5 |
| Gemini 3.8 Flash         | 91.3    | 88.2   | 76.1              | 72.7 | 72.3 |
| GPT-6 luna               | 93.2    | 90.3   | 76.4              | 63.5 | 62.4 |
| Gemini 2.5 Flash         | 94.3    | 92.1   | 76.6              | 66.2 | 66.4 |
| Stanza (matched)         | 96.0    | 95.7   | 86.3              | 72.8 | 69.6 |
| UDPipe 2 (matched)       | 96.5    | 96.1   | 87.2              | 77.0 | 73.7 |
| GREEK grc-ptnk           |         |        |                   |      |      |
| Claude Opus 5.5          | 96.9    | 94.0   | 85.8              | 87.9 | 83.0 |
| Gemini 3.8 Flash         | 90.6    | 87.7   | 80.5              | 83.1 | 78.4 |
| GPT-6 luna               | 96.8    | 93.9   | 84.5              | 83.2 | 77.7 |
| Gemini 2.5 Flash         | 96.2    | 92.3   | 84.0              | 83.2 | 78.3 |
| Stanza (matched)         | 98.6    | 98.3   | 91.7              | 86.0 | 79.4 |
| UDPipe 2 (matched)       | 98.6    | 98.1   | 88.9              | 88.9 | 83.6 |

(b) The AGDT sentences against my conversion and the official UD Perseus trees

|                    | Greek    |        |        | Latin    |        |        |
|--------------------|----------|--------|--------|----------|--------|--------|
|                    | UAS mine | UAS UD | LAS UD | UAS mine | UAS UD | LAS UD |
| Claude Opus 5.5    | 89.1     | 85.9   | 68.6   | 92.0     | 85.4   | 73.2   |
| Gemini 3.8 Flash   | 88.2     | 85.6   | 68.9   | 90.7     | 85.1   | 73.2   |
| GPT-6 luna         | 83.3     | 79.8   | 62.5   | 78.7     | 74.2   | 62.4   |
| Gemini 2.5 Flash   | 81.8     | 78.7   | 62.2   | 80.4     | 74.2   | 64.7   |
| Stanza (matched)   | 74.5     | 76.9   | 70.2   | 66.4     | 63.7   | 50.5   |
| UDPipe 2 (matched) | 76.4     | 80.0   | 73.8   | 69.8     | 64.7   | 55.9   |

TABLE 3. Without lenient conventions the parsers lead on exact morphology, and against the official trees the models' lead on heads shrinks but stays. Features (exact) requires all seven features of a word to match. CLAS scores content words. Latin has 27 sentences aligned with the official trees, Greek 170.

### 5.4 Price

The models span almost two orders of magnitude in price (Figure 8). GPT-6 luna costs 0.10 cents per Latin sentence, and Opus 53 times as much. Gemini 3.8 Flash costs 44 times as much as luna in Latin and is the slowest, at 102.0 seconds per Latin sentence, because through the Gateway it reasons for thousands of tokens even at its minimal setting. Called directly, as Interlinea calls it in production, it reasoned far less on tutor and context prompts, so these figures are an upper bound for Gemini. Because a library sentence is analysed once and then served from the store, the best model’s price is paid once per sentence, not once per reader.

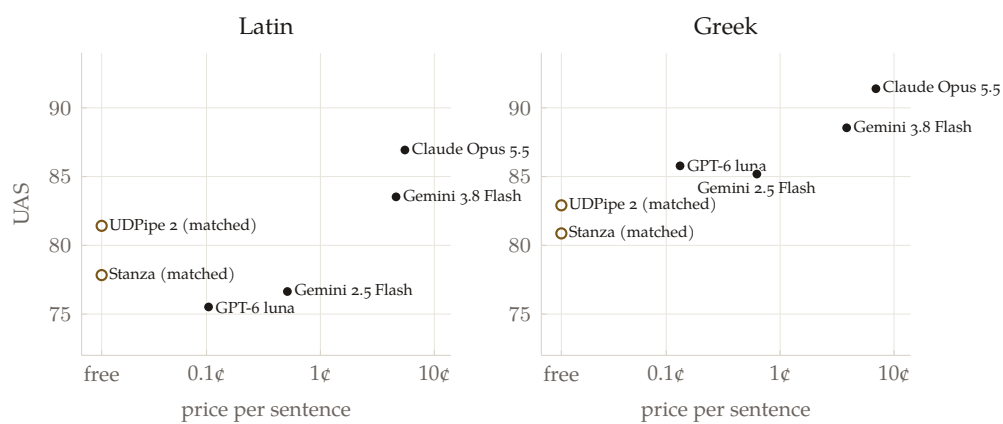


FIGURE 8. Opus buys its lead over the cheapest model at about fifty times the price. Price per sentence against UAS on the public sets, log scale.

## 6 ANALYSIS

### 6.1 Numbered words

When GPT-6 luna gets the bare sentence and must split and number the words itself, its lemmas and features are unchanged but its UAS falls by 15.6 points (Figure 10). Segmentation explains little of it. Its words differ in number from the gold’s in 5.4% of sentences, where it reorders verse into prose order or drops a word, twice a *non*. On the 458 sentences it splits exactly as the gold does, its UAS is still 68.2 against 82.5. Figure 9 shows why. Heads are indices, and a model that numbers its own words points 11.0% of them one word off, against 3.7% with numbered words.

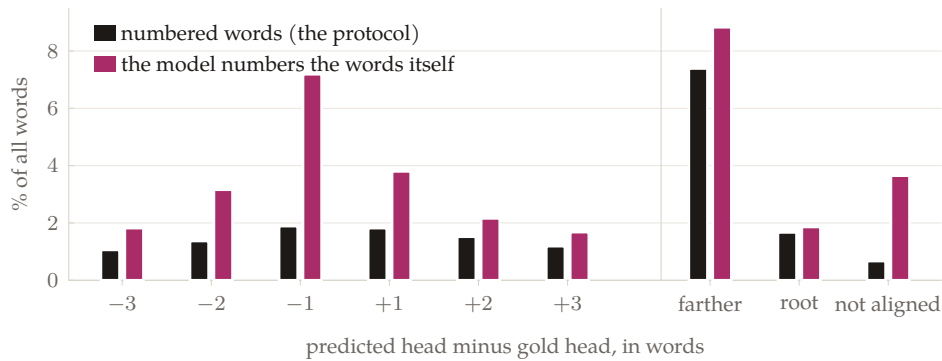


FIGURE 9. A model that numbers the words itself points its heads at neighbours. Where GPT-6 luna’s wrong heads point, as a share of all words, with numbered words and when it splits and numbers the sentence itself.

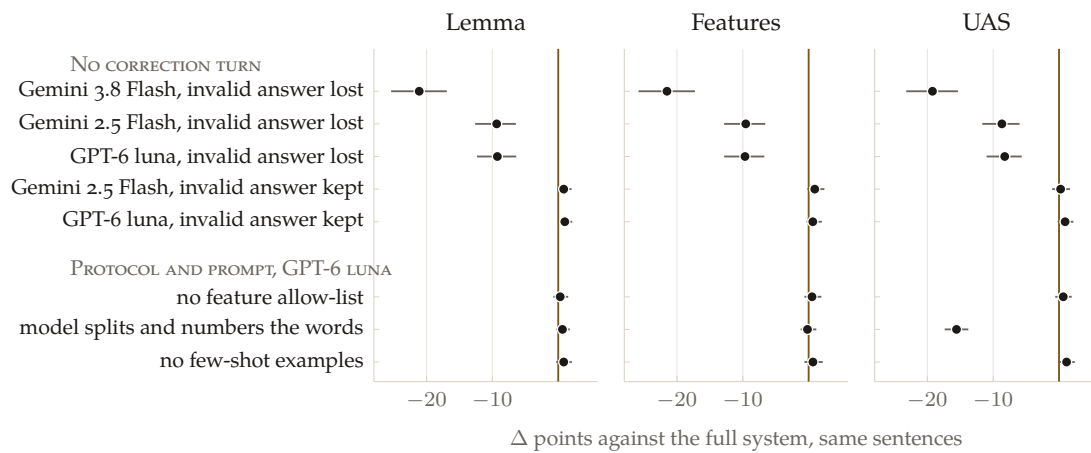


FIGURE 10. Numbering the words is worth 15.6 UAS points, and the correction turn’s worth is validity, not accuracy. Each row is a variant minus the full system on the same sentences, with 95% paired intervals. “Lost” scores a failed first answer as invalid, “kept” scores it as it stands. Opus never needed the turn.

## 6.2 The correction turn

If a sentence whose first answer failed were simply lost, luna would lose 9.2 lemma points and Gemini 3.8 Flash 21.1 (Figure 10). The turn leaves a valid answer for about nine in ten of those sentences. But a failed first answer, scored as it stands, is as accurate word by word as the repaired one, for luna and for Gemini 2.5 Flash alike. What the validator catches is form, a second root, a cycle, or Case=Fem. With the turn, 99.6% of luna's answers are valid, against 93.6% without it, and the reader never sees a malformed one. The examples and the allow-list in the prompt change nothing measurable.

Every one of Gemini 3.8 Flash's 130 failures was an answer cut off before its JSON closed (Figure 11). Its hidden reasoning, heavy through the Gateway, counts against the 16,000-token cap, and the first answers that closed pile up just below it (Figure 12). Rerun once at 32,000 tokens, 119 of the 130 came back valid, for \$9.50 against \$13.19 for the original two calls.

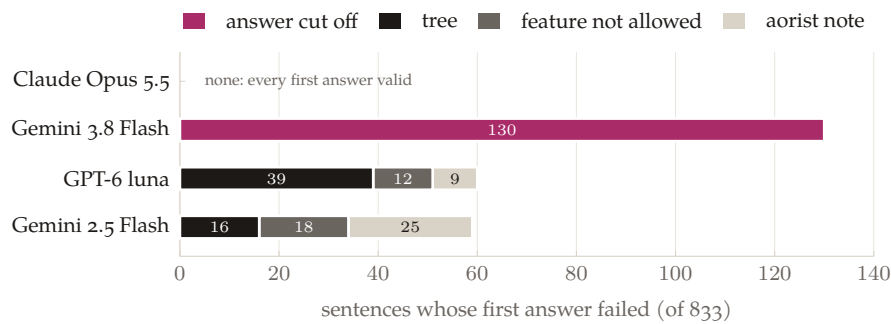


FIGURE 11. Gemini 3.8 Flash's first answers fail because they are cut off, the others' because the validator catches an error. Sentences whose first answer failed, by the first message that applies.

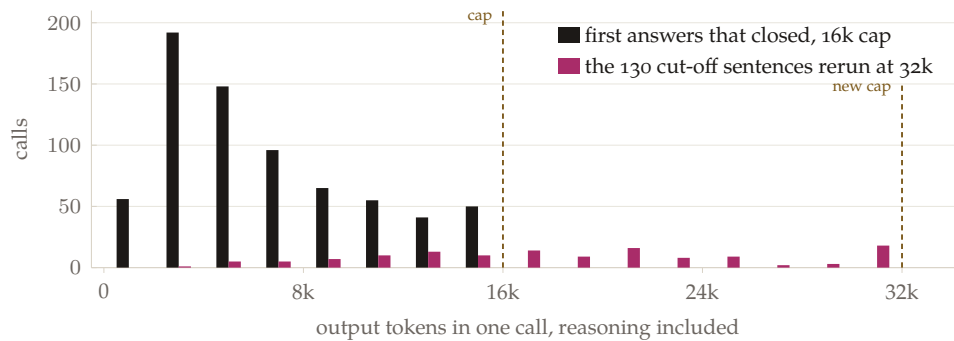


FIGURE 12. Gemini 3.8 Flash's reasoning counts against its output cap. No first answer that closed used more than 15,984 tokens, and with a 32k cap the cut-off sentences spread up to the new limit.

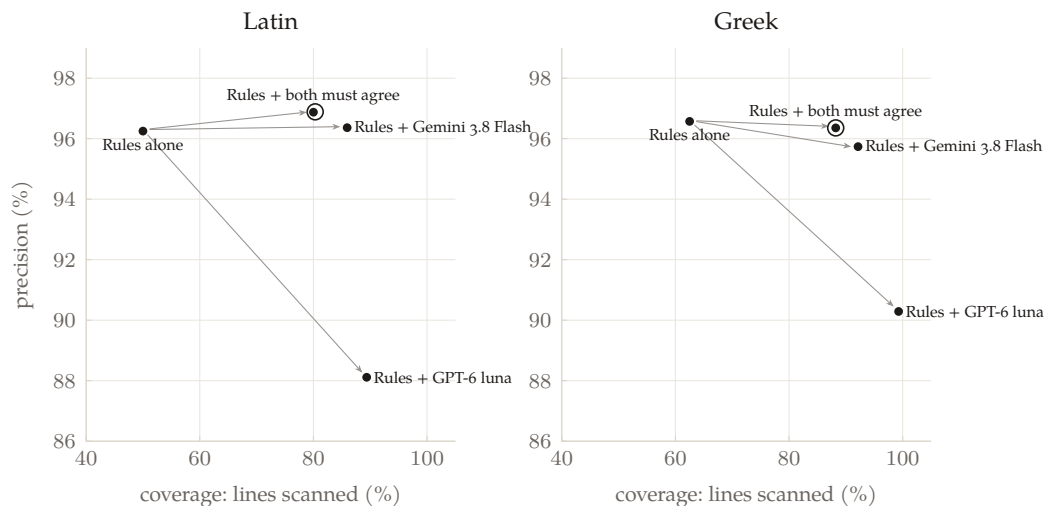


FIGURE 13. Model picks among the rules' candidates raise coverage at little cost in precision. Share of 600 Hypotactic lines scanned against the share scanned correctly. The hollow point is Interlinea's rule.

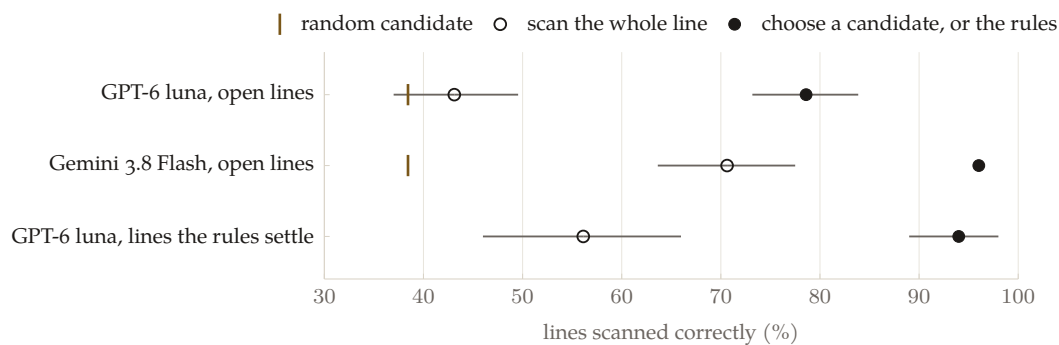


FIGURE 14. Choosing is far easier than scanning. Lines scanned correctly when a model chooses among the rules' candidates, and when it scans the whole line, with 95% intervals. Gemini's whole-line figure counts only the 160 lines it answered in time.

### 6.3 Scansion

I took 600 lines from fifteen works scanned in Hypotactic (Chamberlain, 2026), from Homer, Theognis, Aeschylus and Euripides to Virgil, Ovid, Catullus, Horace, Phaedrus and Seneca. The rules alone scan 55.8% of them at 96.4% precision. With Gemini 3.8 Flash and GPT-6 luna agreeing on the rest, the system scans 83.8% at 96.6% (Figure 13). The candidates do most of the work (Figure 14). Luna picks the right one 78.6% of the time, but scans the same lines whole at 43.1%, and on lines the rules settle alone it reaches 56.1% against the rules' 94.0%. Agreement adds little precision over Gemini alone (+1.6 points,  $p = 0.087$ ) and mainly buys abstention. Results by metre are in Appendix D.

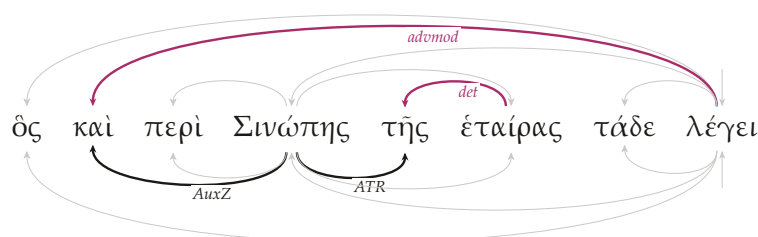
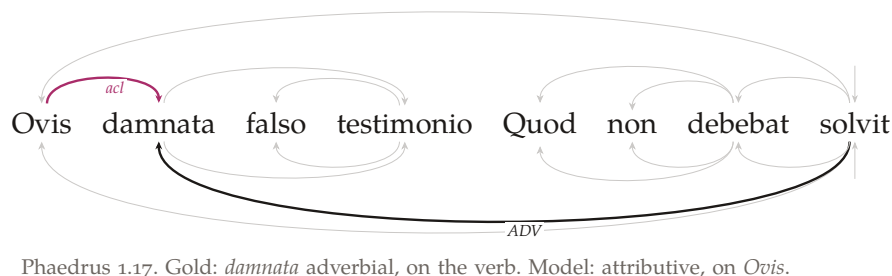
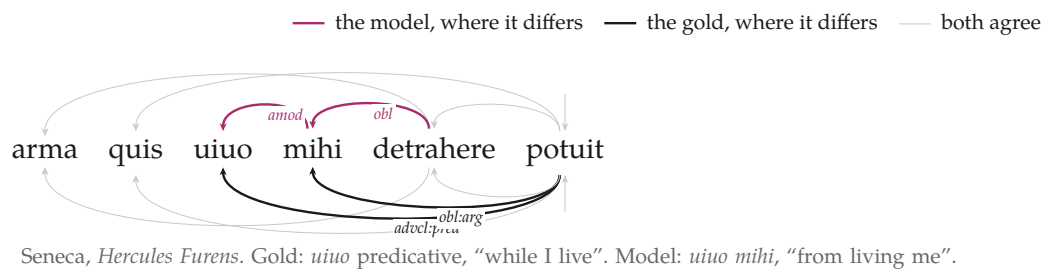


FIGURE 15. Some of the best model’s attachment errors are defensible readings. Three of Claude Opus 5.5’s disagreements with the gold, the model’s tree above the words and the gold’s below, in colour where they differ.

#### 6.4 What the disagreements look like

In the Seneca line of Figure 15, Opus takes *uiuo mihi* as “from me, alive”, where the gold makes *uiuo* predicative. In Phaedrus it reads *damnata* as an attribute of *Ovis* where the gold makes it adverbial, the old ambiguity of the participle. A single gold standard scores a choice between schools of annotation as an error. Case errors show the same split (Figure 16). Opus’s largest Latin confusion, accusative read as ablative, comes entirely from the Lucca charters, where scribes wrote *uno capo* where the syntax needs an accusative. The treebank annotates the function, and the model reads the form.

## 7 CONCLUSION

Language models annotate Latin and Greek well enough to read with if code owns the form of their answers. Owning the form is worth more than I expected. Numbered words are worth 15.6 UAS points, because a head is an index and a model that numbers its own words points to the wrong one. The validator does not make a model right, but it makes malformed answers visible, and one correction turn makes nine in ten of them valid. The best model finds heads better than parsers trained on the test treebanks, and the parsers keep exact morphology and relation labels. In verse, a model asked to choose among a few rule-built readings is far more accurate than one asked to scan.

## LIMITATIONS

Each gold set comes from one annotator or project, and three public sets were converted between schemes, by their maintainers or by me. My AGDT conversion favours the models by several UAS points against the official trees, and only 27 Latin Perseus sentences align with them. The two sets released after the cutoffs are small and differ in genre. The parsers ran with default settings on gold words, and stronger Latin parsers exist (Straka et al., 2024). Glosses, notes and translations, which readers use most, are not scored, and nothing here measures whether reading with Interlinea teaches anyone Latin or Greek. The models are commercial services that change, and each was tested once, in October 2026.

## REPRODUCIBILITY

The gold files, every model answer and the per-sentence scores are available from the author, so every number can be recomputed without calling a model. Appendix A outlines the prompts. Interlinea’s source code and the prompts’ full text are not public.

## ACKNOWLEDGEMENTS

I thank the makers of the treebanks and scansions on which this evaluation rests, and David Chamberlain for releasing Hypotactic’s data under an open licence.

## REFERENCES

- Alpheios. Alpheios: Reading tools for classical languages and literatures. The Alpheios Project, Ltd., 2026. URL <https://alpheios.net/>. Open-source reader, browser extension and embeddable library for Latin, Greek, Arabic and Persian. Accessed 20 October 2026.
- Anthropic. System card: Claude Opus 5.5. Anthropic, September 2026. URL <https://www.anthropic.com/claude-opus-5-5-system-card>. Published 22 September 2026. Section 1.1 states a knowledge cutoff of June 2026; the Claude Platform model overview lists reliable-knowledge and training-data cutoffs of June 2026.
- David Bamman and Patrick J. Burns. Latin BERT: A contextual language model for classical philology, 2020. URL <https://arxiv.org/abs/2009.10053>.
- David Bamman and Gregory Crane. The Ancient Greek and Latin dependency treebanks. In Caroline Sporleder, Antal van den Bosch, and Kalliope Zervanou, editors, *Language Technology*

- for *Cultural Heritage*, pages 79–98. Springer, Berlin, Heidelberg, 2011. doi: 10.1007/978-3-642-20227-8\_5.
- Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. An empirical investigation of statistical significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL <https://aclanthology.org/D12-1091/>.
- Flavio Massimiliano Cecchini, Timo Korkiakangas, and Marco Passarotti. A new Latin treebank for Universal Dependencies: Charters between Ancient Latin and Romance languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 933–942, Marseille, France, May 2020a. European Language Resources Association. URL <https://aclanthology.org/2020.lrec-1.117/>.
- Flavio Massimiliano Cecchini, Rachele Sprugnoli, Giovanni Moretti, and Marco Passarotti. UDante: First steps towards the Universal Dependencies treebank of Dante’s Latin works. In Johanna Monti, Felice Dell’Orletta, and Fabio Tamburini, editors, *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)*, volume 2769 of *CEUR Workshop Proceedings*, Bologna, Italy, 2020b. CEUR-WS.org. URL [https://ceur-ws.org/Vol-2769/paper\\_14.pdf](https://ceur-ws.org/Vol-2769/paper_14.pdf).
- David Chamberlain. Hypotactic: Greek and Roman verse, audio-visually enhanced, 2026. URL <https://hypotactic.com/>. Accessed 20 October 2026.
- Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*, December 2024a. ISSN 2835-8856. URL <https://openreview.net/forum?id=cSimKw5p6R>. Featured Certification.
- Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. Teaching large language models to self-debug. In *The Twelfth International Conference on Learning Representations (ICLR)*, pages 8746–8825, 2024b. URL [https://proceedings.iclr.cc/paper\\_files/paper/2024/hash/2460396f2d0d421885997dd1612ac56b-Abstract-Conference.html](https://proceedings.iclr.cc/paper_files/paper/2024/hash/2460396f2d0d421885997dd1612ac56b-Abstract-Conference.html).
- CLTK. Scansion of poetry: HexameterScanner, PentameterScanner and HendecasyllableScanner. The Classical Language Toolkit, legacy (vo.1) documentation, Latin module, 2019. URL <https://v0.cltk.org/en/latest/latin.html#scansion-of-poetry>. Documentation copyright 2019 Kyle P. Johnson. Accessed 20 October 2026.
- Emanuela Colombi, Luca Mondin, Luigi Tassarolo, and Andrea Bacianini. Pedecerto: Metrica latina digitale. Università degli studi di Udine, 2011. URL <https://www.pedecerto.eu/>. Accessed 20 October 2026.
- Gregory Crane. Generating and parsing classical Greek. *Literary and Linguistic Computing*, 6(4): 243–245, 1991. doi: 10.1093/lc/6.4.243.
- Eric Cullhed and Albin Thörn Cleland. Stoicheia: Character-level masked diffusion for Ancient Greek textual restoration, parsing, and metrical scansion, 2026. URL <https://arxiv.org/abs/2608.07249>.
- Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker’s guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1128.
- Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Number 57 in Monographs on Statistics and Applied Probability. Chapman & Hall, New York, 1993. ISBN 0412042312. doi: 10.1201/9780429246593.
- Anise Ferreira. UNESP trees: Coleção de treebanks dos Projetos Abertos em Letras Clássicas

- Digitais. Perseids Publications, GitHub repository, 2021. URL <https://github.com/perseids-publications/unesp-trees>. Faculdade de Ciências e Letras, UNESP Araraquara.
- Saibo Geng, Martin Josifoski, Maxime Peyrard, and Robert West. Grammar-constrained decoding for structured NLP tasks without finetuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10932–10952, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.674.
- Google DeepMind. Gemini 3.8 Flash model card. Google DeepMind, September 2026. URL <https://deepmind.google/models/model-cards/gemini-3-8-flash/>. Published 2 September 2026. States a knowledge cutoff of March 2026, adding that in some domains knowledge may be limited to January 2025, in line with the Gemini 3 model family.
- Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *The Twelfth International Conference on Learning Representations (ICLR)*, pages 57734–57811, 2024. URL [https://proceedings.iclr.cc/paper\\_files/paper/2024/hash/fe126561bbf9d4467dbb8d27334b8fe-Abstract-Conference.html](https://proceedings.iclr.cc/paper_files/paper/2024/hash/fe126561bbf9d4467dbb8d27334b8fe-Abstract-Conference.html).
- Dag T. T. Haug and Marius L. Jøhndal. Creating a parallel treebank of the Old Indo-European Bible translations. In Caroline Sporleder and Kiril Ribarov, editors, *Proceedings of the LREC 2008 Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, pages 27–34, 2008. URL [http://www.lrec-conf.org/proceedings/lrec2008/workshops/W22\\_Proceedings.pdf](http://www.lrec-conf.org/proceedings/lrec2008/workshops/W22_Proceedings.pdf).
- Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL <https://www.jstor.org/stable/4615733>.
- Claudiu Daniel Hromei, Danilo Croce, and Roberto Basili. U-DepPLLaMA: Universal dependency parsing via auto-regressive large language models. *Italian Journal of Computational Linguistics (IJCoL)*, 10(1):21–38, 2024. doi: 10.4000/125nm.
- Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations (ICLR)*, pages 32808–32824, 2024. URL [https://proceedings.iclr.cc/paper\\_files/paper/2024/hash/8b4add8b0aa8749d80a34ca5d941c355-Abstract-Conference.html](https://proceedings.iclr.cc/paper_files/paper/2024/hash/8b4add8b0aa8749d80a34ca5d941c355-Abstract-Conference.html).
- Federica Iurescia, Federica Gamba, Flavio Massimiliano Cecchini, Francesco Mambrini, Giovanni Moretti, Marco Passarotti, and Paolo Ruffolo. UD\_Latin-CIRCSE. Universal Dependencies treebank, first released in UD v2.14, 2024. URL [https://github.com/UniversalDependencies/UD\\_Latin-CIRCSE](https://github.com/UniversalDependencies/UD_Latin-CIRCSE). Contributors as listed in the treebank README; the README names no publication to cite.
- Federica Iurescia, Marco Passarotti, and Rachele Sprugnoli. Overview of the dependency parsing task at EvaLatin 2026. In *Proceedings of the Fourth Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2026) @ LREC 2026*, pages 211–218, Palma, Mallorca (Spain), May 2026. ELRA Language Resources Association (ELRA). doi: 10.63317/2j3y8jv8ktyd.
- Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. When can LLMs actually correct their own mistakes? a critical survey of self-correction of LLMs. *Transactions of the Association for Computational Linguistics*, 12:1417–1440, 2024. doi: 10.1162/tacl\_a\_00713.
- Alek Keersmaekers. The GLAUx corpus: methodological issues in designing a long-term, diverse, multi-layered corpus of Ancient Greek. In *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021*, pages 39–50, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.lchange-1.6.
- Jason Liu and Ivan Leo. Instructor: Structured outputs for LLMs, 2026. URL <https://github.com/567-labs/instructor>. Python library; Pydantic validation with automatic retries; MIT licence.

- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative refinement with self-feedback. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc., 2023. doi: 10.52202/075280-2019.
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France, May 2020. European Language Resources Association. URL <https://aclanthology.org/2020.lrec-1.497/>.
- OpenAI. GPT-6 Luna model. OpenAI API documentation, 2026. URL <https://developers.openai.com/api/docs/models/gpt-6-luna>. States a knowledge cutoff of 18 May 2026. Accessed 20 October 2026.
- Marco Passarotti. The project of the Index Thomisticus treebank. In Monica Berti, editor, *Digital Classical Philology*, pages 299–320. De Gruyter, 2019. doi: 10.1515/9783110599572-017.
- Marco Passarotti, Marco Budassi, Eleonora Litta, and Paolo Ruffolo. The Lemlat 3.0 package for morphological analysis of Latin. In Gerlof Bouma and Yvonne Adesam, editors, *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, pages 24–31, Gothenburg, May 2017. Linköping University Electronic Press. URL <https://aclanthology.org/W17-0506/>.
- Gabriel H. Pizzorno. Correction as annotation: Bootstrapping a dependency parser for documentary medieval Latin, 2026. URL <https://arxiv.org/abs/2609.06266>.
- Luc Pommeret, Thibault Wagret, and Jules Deret. THIVLVC: Retrieval augmented dependency parsing for Latin. In *Proceedings of the Fourth Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2026) @ LREC 2026*, pages 219–225, Palma, Mallorca (Spain), May 2026. ELRA Language Resources Association (ELRA). doi: 10.63317/2q8twojtotyb.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-demos.14.
- D. S. Raven. *Latin Metre: An Introduction*. Faber and Faber, London, 1965.
- Frederick Riemenschneider and Anette Frank. Exploring large language models for classical philology. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15181–15199, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.846.
- Manley Roberts, Himanshu Thakur, Christine Herlihy, Colin White, and Samuel Dooley. To the cutoff... and beyond? A longitudinal perspective on LLM data contamination. In *The Twelfth International Conference on Learning Representations (ICLR)*, pages 10795–10862, 2024. URL [https://proceedings.iclr.cc/paper\\_files/paper/2024/hash/2d04d97593c8c33d415337f408ed0e1b-Abstract-Conference.html](https://proceedings.iclr.cc/paper_files/paper/2024/hash/2d04d97593c8c33d415337f408ed0e1b-Abstract-Conference.html).
- Pablo Ruiz Fabo, Anxo Alonso Pérez, Pablo Rodríguez Fernández, and Pablo Gamallo. Automatic metrical scansion of poetry in a low-resource setting. In *Proceedings of Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026*, pages 114–125, Palma de Mallorca (Spain), May 2026. ELRA Language Resources Association (ELRA). doi: 10.63317/4pzd6u7388jm.
- Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. NLP evaluation in trouble: On the need to measure LLM data contamination for each

- benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.722.
- Anne-Kathrin Schumann, Christoph Beierle, and Norbert Blößner. Using finite-state machines to automatically scan classical Greek hexameter, 2021. URL <https://arxiv.org/abs/2101.11437>.
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 8634–8652. Curran Associates, Inc., 2023. doi: 10.52202/075280-0377.
- Rachele Sprugnoli, Marco Passarotti, Flavio Massimiliano Cecchini, and Matteo Pellegrini. Overview of the EvaLatin 2020 evaluation campaign. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 105–110, Marseille, France, May 2020. European Language Resources Association (ELRA). URL <https://aclanthology.org/2020.lt4hala-1.16/>.
- Rachele Sprugnoli, Marco Passarotti, Flavio Massimiliano Cecchini, Margherita Fantoli, and Giovanni Moretti. Overview of the EvaLatin 2022 evaluation campaign. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 183–188, Marseille, France, June 2022. European Language Resources Association. URL <https://aclanthology.org/2022.lt4hala-1.29/>.
- Rachele Sprugnoli, Federica Iurescia, and Marco Passarotti. Overview of the EvaLatin 2024 evaluation campaign. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 190–197, Torino, Italia, May 2024. ELRA and ICCL. URL <https://aclanthology.org/2024.lt4hala-1.21/>.
- Milan Straka. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/K18-2020.
- Milan Straka, Jana Straková, and Federica Gamba. ÚFAL LatinPipe at EvaLatin 2024: Morphosyntactic analysis of Latin. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 207–214, Torino, Italia, May 2024. ELRA and ICCL. URL <https://aclanthology.org/2024.lt4hala-1.24/>.
- Elina Stüssi and Phillip Ströbel. Part-of-speech tagging of 16th-century Latin with GPT. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 196–206, St. Julians, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.latechclfl.1.18.
- Colin Swaelens, Ilse De Vos, and Els Lefever. Lemmatisation & morphological analysis of unedited Greek: Do simple tasks need complex solutions? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7681–7689, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.399.
- Daniel G. Swanson, Bryce D. Bussert, and Francis Tyers. Producing a parallel Universal Dependencies treebank of Ancient Hebrew and Ancient Greek via cross-lingual projection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13074–13078, Torino, Italia, May 2024. ELRA and ICCL. URL <https://aclanthology.org/2024.lrec-main.1145/>.
- Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. Let me speak freely? a study on the impact of format restrictions on large language model performance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1218–1236, Miami, Florida, US, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-industry.91.

- Albin Thörn Cleland and Eric Cullhed. Automatic annotation of Ancient Greek vowel length, 2026. URL <https://arxiv.org/abs/2608.01935>.
- Chahan Vidal-Gorène, Bastien Kindt, and Florian Cafiero. Under-resourced studies of under-resourced languages: Lemmatization and POS-tagging with LLM annotators for historical Armenian, Georgian, Greek and Syriac, 2026. URL <https://arxiv.org/abs/2602.15753>.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL <https://openreview.net/forum?id=1PL1NIMMrw>.
- M. L. West. *Greek Metre*. Clarendon Press, Oxford, 1982. ISBN 0198140185.
- Brandon T. Willard and Rémi Louf. Efficient guided generation for large language models, 2023. URL <https://arxiv.org/abs/2307.09702>.
- Daniel Zeman et al. Universal Dependencies 2.18. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Charles University, May 2026. URL <http://hdl.handle.net/11234/1-6149>. Released 15 May 2026.

## A THE PROMPTS, IN OUTLINE

The prompts’ full text is not published. In outline:

- **Annotation.** A system message sets the task and the output contract: one object per numbered word, with lemma, dictionary entry, UD part of speech, UD features, head, relation, gloss, a learner’s note and an enclitic slot, and a translation of the sentence. It states the UD conventions Interlinea follows where treebanks differ, such as writing the Greek aorist as Tense=Past|Aspect=Perf in every mood, names mistakes models made in development, such as Number=3 for a third person, and lists the feature values the language allows. Two hand-checked example sentences per language follow as earlier turns of the conversation. The user turn gives the language, the sentence and its numbered words.
- **Correction turn.** The model’s previous answer is returned with the validator’s messages, one line per failed condition (for example token 7: a word cannot be its own head), and the model is asked to return the whole annotation again with those problems fixed.
- **Scansion.** The model is told the metre and its template, sees each open line with its numbered candidate scansions, and is asked for one option number per line, using the true vowel lengths of the words.

## B SCORING CONVENTIONS

- **Lemma.** Compared after Unicode decomposition with all combining marks removed, lower-cased, final sigma normalized, and *j* and *v* read as *i* and *u*.
- **Part of speech.** PRON and DET are interchangeable, as are PROPON and NOUN, NUM and ADJ, and AUX and VERB for *sum*, *εἰμί* and *possum*. Negations may be PART or ADV; connectives and discourse particles, Latin and Greek, may be CCONJ, PART or ADV. Pronominal adjectives (possessives, quantifiers, *alter*, *multus*) may be ADJ or DET; *nihil* and *nemo* may be NOUN or PRON.
- **Features.** A multi-valued gold feature (Voice=Mid, Pass) accepts any of its values. The Greek perfect accepts Tense=Pres or Past. A prediction with Aspect but no Tense is read as Perf → Past, Imp → Pres, Prosp → Fut. The copula's voice is not scored.
- **Heads.** AGDT trees are re-rooted to UD style, so that prepositions, subordinators, coordinators and the copula depend on their content word. A head that attaches to punctuation or cannot be mapped is not scored.

## C RESULTS BY GOLD SET

| Set                             | System             | Lemma | UPOS | Feats | UAS  | LAS  |
|---------------------------------|--------------------|-------|------|-------|------|------|
| UD Latin CIRCSE<br>(test)       | Claude Opus 5.5    | 94.3  | 96.2 | 98.8  | 88.1 | 78.9 |
|                                 | Gemini 3.8 Flash   | 89.9  | 91.5 | 95.5  | 83.8 | 76.0 |
|                                 | GPT-6 luna         | 86.9  | 90.9 | 91.8  | 73.7 | 65.8 |
|                                 | Gemini 2.5 Flash   | 89.4  | 92.4 | 93.7  | 75.6 | 67.8 |
|                                 | Stanza (matched)   | 88.2  | 93.0 | 89.4  | 64.8 | 55.0 |
|                                 | UDPipe 2 (matched) | 90.6  | 93.8 | 89.3  | 70.8 | 62.9 |
| UD Latin LLCT (test)            | Claude Opus 5.5    | 91.6  | 96.5 | 91.2  | 81.8 | 72.0 |
|                                 | Gemini 3.8 Flash   | 86.9  | 91.1 | 87.0  | 77.7 | 69.0 |
|                                 | GPT-6 luna         | 89.5  | 95.7 | 90.3  | 72.7 | 61.0 |
|                                 | Gemini 2.5 Flash   | 90.9  | 96.4 | 89.6  | 74.0 | 64.5 |
|                                 | Stanza (matched)   | 97.1  | 99.2 | 97.4  | 94.1 | 92.2 |
|                                 | UDPipe 2 (matched) | 96.5  | 99.3 | 97.5  | 94.4 | 92.4 |
| AGDT Latin (UD<br>Perseus test) | Claude Opus 5.5    | 90.6  | 93.7 | 96.0  | 93.0 | –    |
|                                 | Gemini 3.8 Flash   | 92.4  | 94.2 | 96.8  | 92.2 | –    |
|                                 | GPT-6 luna         | 88.6  | 93.1 | 95.7  | 83.0 | –    |
|                                 | Gemini 2.5 Flash   | 88.9  | 93.9 | 95.1  | 82.5 | –    |
|                                 | Stanza (matched)   | 79.1  | 88.6 | 89.5  | 74.7 | –    |
|                                 | UDPipe 2 (matched) | 87.7  | 89.5 | 90.3  | 79.1 | –    |
| AGDT Greek (UD<br>Perseus test) | Claude Opus 5.5    | 93.9  | 93.8 | 97.7  | 89.1 | –    |
|                                 | Gemini 3.8 Flash   | 93.1  | 93.1 | 96.6  | 88.2 | –    |
|                                 | GPT-6 luna         | 92.8  | 94.1 | 96.9  | 83.3 | –    |
|                                 | Gemini 2.5 Flash   | 90.7  | 92.3 | 94.2  | 81.8 | –    |
|                                 | Stanza (matched)   | 87.1  | 92.9 | 95.1  | 74.5 | –    |
|                                 | UDPipe 2 (matched) | 88.4  | 94.2 | 95.5  | 76.4 | –    |
| UD Greek PTNK<br>(test)         | Claude Opus 5.5    | 95.2  | 96.9 | 97.1  | 95.3 | 87.9 |
|                                 | Gemini 3.8 Flash   | 89.6  | 90.6 | 90.9  | 89.2 | 83.1 |
|                                 | GPT-6 luna         | 95.2  | 96.8 | 97.2  | 90.1 | 83.2 |
|                                 | Gemini 2.5 Flash   | 95.2  | 96.2 | 96.6  | 91.0 | 83.2 |
|                                 | Stanza (matched)   | 94.7  | 98.6 | 93.7  | 91.9 | 86.0 |
|                                 | UDPipe 2 (matched) | 95.3  | 98.6 | 93.5  | 94.2 | 88.9 |
| DALME-Marseille<br>gold         | Claude Opus 5.5    | 82.9  | 92.7 | 77.8  | 93.8 | 81.4 |
|                                 | Gemini 3.8 Flash   | 83.6  | 92.6 | 80.5  | 93.7 | 80.6 |
|                                 | GPT-6 luna         | 82.1  | 85.1 | 74.7  | 84.0 | 65.3 |
|                                 | Gemini 2.5 Flash   | 83.5  | 89.5 | 78.3  | 85.7 | 69.8 |
|                                 | Stanza (matched)   | 69.3  | 74.7 | 69.8  | 74.7 | 52.2 |
|                                 | UDPipe 2 (matched) | 67.8  | 75.4 | 67.6  | 76.8 | 51.3 |
| UNESP Artemidorus               | Claude Opus 5.5    | 90.2  | 93.7 | 94.2  | 70.5 | –    |
|                                 | Gemini 3.8 Flash   | 86.1  | 89.0 | 90.2  | 67.5 | –    |
|                                 | GPT-6 luna         | 84.3  | 88.4 | 88.9  | 63.1 | –    |
|                                 | Gemini 2.5 Flash   | 88.3  | 92.5 | 91.8  | 65.9 | –    |
|                                 | Stanza (matched)   | 92.5  | 94.7 | 96.4  | 74.7 | –    |
|                                 | UDPipe 2 (matched) | 90.0  | 94.3 | 94.5  | 73.8 | –    |

TABLE 4. Results on each gold set, in percent. LAS is on UD-native sets only. For the two sets released in August 2026, the parsers use the closest treebank, LLCT for DALME-Marseille and Perseus for Artemidorus.

## D SCANSION BY METRE

| Metre                    | Lines | rules alone |           | Interlinea's rule |           |
|--------------------------|-------|-------------|-----------|-------------------|-----------|
|                          |       | coverage    | precision | coverage          | precision |
| Greek dactylic hexameter | 80    | 98.8        | 94.9      | 98.8              | 94.9      |
| Greek elegiac couplet    | 40    | 97.5        | 100.0     | 100.0             | 100.0     |
| Greek iambic trimeter    | 160   | 35.6        | 96.5      | 80.0              | 96.1      |
| Latin choliambic         | 4     | 25.0        | 100.0     | 100.0             | 100.0     |
| Latin dactylic hexameter | 91    | 70.3        | 95.3      | 95.6              | 95.4      |
| Latin elegiac couplet    | 54    | 88.9        | 97.9      | 100.0             | 98.1      |
| Latin greater Asclepiad  | 3     | 100.0       | 100.0     | 100.0             | 100.0     |
| Latin hendecasyllable    | 10    | 20.0        | 100.0     | 70.0              | 100.0     |
| Latin iambic senarius    | 40    | 0.0         | –         | 27.5              | 100.0     |
| Latin iambic trimeter    | 80    | 20.0        | 93.8      | 72.5              | 96.6      |
| Latin Sapphic stanza     | 38    | 68.4        | 96.2      | 86.8              | 97.0      |

TABLE 5. Scansion of the 600 Hypotactic lines by metre, in percent. Interlinea's rule trusts Gemini 3.8 Flash alone in dactylic metres and requires GPT-6 luna to agree elsewhere. The metre and a line's place in its stanza come from the gold.

## E THREE FAILURES AND THEIR REPAIRS

| SENTENCE                                                                                                  | THE VALIDATOR, VERBATIM                                                       | AFTER THE CORRECTION TURN                                                                                           |
|-----------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------|
| <i>nunc Hercule opus est perfer<br/>hanc molem mali</i><br>Seneca, <i>Hercules Furens</i> , GPT-6<br>luna | token 3: heads form a cycle<br>token 4: heads form a cycle                    | A valid tree, rooted on <i>perfer</i> .<br>The gold roots it on <i>opus</i> , so<br>valid is not the same as right. |
| <i>Tantum effata, furens antro<br/>se immisit aperto</i><br>Virgil, <i>Aeneid</i> 6, GPT-6 luna           | token 2: Case=Fem is not allowed (use Nom,<br>Gen, Dat, Acc, Abl, Voc, Loc)   | The gender moves to<br>Gender=Fem, but the case is<br>dropped, where the gold has<br>Case=Nom.                      |
| <i>καὶ τυχῶν ἤκεν εἰς<br/>Κρότωνα.</i><br>Athenaeus, <i>Deipnosophistae</i> ,<br>GPT-6 luna               | token 3: the note calls it an aorist, so<br>feats need Tense=Past Aspect=Perf | Note and features now agree,<br>and the tense matches the<br>gold.                                                  |

TABLE 6. Three of GPT-6 luna's first-answer failures. The first answer is not stored, and the validator's messages describe it.

## F CASE CONFUSIONS



FIGURE 16. Gold case against predicted case on the public sets, as a percentage of each gold case, for Claude Opus 5.5 and Stanza (matched). *None* means no case was given, including every word of a sentence whose answer never validated.